Statistical Inference for High Dimensional Problems

Statistical Inference for High Dimensional Problems

A dissertation presented

by

Rajarshi Mukherjee

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2014

Dissertation Advisor: Professor Xihong Lin Rajarshi Mukherjee

Statistical Inference for High Dimensional Problems

Abstract

In this dissertation, we study minimax hypothesis testing in high-dimensional regression against sparse alternatives and minimax estimation of average treatment effect in an semiparametric regression with possibly large number of covariates.

In Chapter 1, we investigate minimax detection boundary of a class of tests based on penalized variable selection procedures when testing for a global null against sparse alternatives. We demonstrate phase transition in performance of these tests based on sparsity of the alternatives and provide comparison of minimax and local power against the generalized likelihood ratio test.

In Chapter 2, we study the detection boundary for minimax hypothesis testing in the context of high-dimensional, sparse binary regression models. Motivated by genetic sequencing association studies for rare variant effects, we investigate the complexity of the hypothesis testing problem when the design matrix also has specific sparsity structures. We observe a new phenomenon in the behavior of detection boundary which does not occur in the case of Gaussian linear regression. We derive the detection boundary as a function of two components: a sparsity interaction parameter between the design matrix and the alternative and the minimal signal strength required for successful detection. If the sparsity interaction parameter of the design matrix is too high, any test is asymptotically powerless irrespective of the magnitude of signal strength. For binary design matrices with not too high sparsity interaction parameter, our results are parallel to the Gaussian case. In this context, we derive detection boundaries for both dense and sparse regimes. For dense regime, our results are rate optimal; for sparse regime, we provide sharp constants. Our optimal tests are extensions of generalized likelihood ratio test and Higher Criticism test.

In Chapter 3, we study estimation of average treatment effect in semiparametric regression using the theory of higher order influence functions under random covariates with no smoothness assumptions on the density of the covariates. We observe a surprising dependence on the orthonormal basis chosen for construction of the estimators. We also characterize relevant third order efficient testing score in a related submodel which might be useful for future research.

Contents

Acknowledgments

I owe my deepest gratitude to my thesis advisor, Professor Xihong Lin, for being a very kind, encouraging and supporting mentor. I am really thankful to Xihong for giving me the support and freedom to work on interesting and challenging problems. It is difficult to express in words my gratitude towards Professor James Robins who has been at the same time a wonderful teacher, selfless mentor and a guardian guiding figure in my graduate life. I especially want to thank Jamie for being patient and supportive with my failed attempts and prolonged unproductive periods in working on a problem from which I learned about a beautiful and challenging area of research. It has been a special privilege to learn from one of the most creative and brilliant minds in the field. I owe special thanks to Dr. Eric Tchetgen Tchetgen for guiding me through various topics in minimax and adaptive inference of nonlinear functionals whenever I struggled. I would like to thank my dissertation committee members, Dr. Tianxi Cai and Dr. Natesh Pillai for their helpful comments and insights and also pointing out which questions are interesting and challenging at the same time. Finally, I would like to thank Dr. Andrea Rotnitzky from the bottom of my heart for being a wonderful and passionate teacher and for introducing me to the world of semiparametric inference and causality through two of the best courses offered during my graduate studies.

I would like to mention special thanks to Matey Nekov, Denis Agniel and Caleb Miles from the Statistics Discussion Group for bearing with my whims of talking about anything remotely related to mathematics, statistics or optimization literature.

This thesis would not have been complete without the continuous support from all my friends in and around Boston. In particular, my heartfelt thanks goes out to Kaustubh Adhikari, Sabyasachi Chatterjee, Dolon Bhattacharya and Sunanda Sanyal for being there for me throughout. I would also like to thank Akash Chattopadhya and Payel Safui for the wonderful board game sessions on cold winter nights. It is difficult to express my thanks in words to my friends from ISI, especially Sayan DasGupta, Joyjit Roy and Anupam Chhibber, whom I whined to and chatted with in the wee hours of night and day.

This dissertation is dedicated to my family. My parents, Anita Mukherjee and Jaydev

Mukherjee and my elder sister Satarupa Mukherjee for being there for me. They never stopped loving me even in the most difficult situations. Wherever I am today and will be tomorrow, I owe it all to them. It is with deepest emotional gratitude I dedicate my dissertation to my grandparents, Nibedita Chakravarti and Jayanta Chakravarti, who are the most influential and loving people in my whole process of growing up as a human being. I owe almost everything in my life to my grandmother Nibedita Chakravarti for believing in me when nobody else did.

Last but not the least, this dissertation would not have been possible without Wan-Chen Lee, to whom I owe everything I cherished, learned and developed in my graduate years. I can never thank her enough for being with me through thick and thin; for being my support when the ground was weak and shaky and for being the magical combination of companion, teacher and listener in the most difficult times of my life - when things looked bleak, gray and unfair. Thank you Chicky!

Hypothesis Testing Using Penalized Likelihoods for High-Dimensional Data

Rajarshi Mukherjee, Xihong Lin and Raymond J. Carroll*

Department of Biostatistics

Harvard School of Public Health

and

* Department of Statistics

Texas A&M University

1.1 Introduction

High dimensional data are commonly observed in health science research, such as genetic and genomic studies. Penalized likelihood is a popular method for performing variable selection for high-dimensional data. A variety of oracle variable selection procedures have been proposed, including LASSO under certain conditions (Meinshausen and Bühlmann, 2006; Meinshausen and Yu, 2009; Zhao and Yu, 2006), SCAD (Fan and Li, 2001; Fan and Peng, 2004; Fan and Lv, 2011), MCP (Zhang, 2010), and SELO (Dicker and Lin, 2012). SCAD and MCP both belong to a class of concave penalized likelihood procedures. These oracle variable selection penalties enjoy the property of consistency of model selection and the oracle property, i.e., by properly choosing the tuning parameter, they estimate the zero components of the true parameter vector exactly as zero with probability approaching one as sample size increases, while still giving consistent estimators of the non-zero components; and the asymptotic distribution of the estimator of the non-zero components is the same as if the true model were known. The recent literature has primarily focused on studying consistency of model selection and asymptotic normality of these penalized likelihood based estimators using the oracle penalties. Limited work has been done on investigating global hypothesis testing using these estimators, especially when both sample size n and the number of variables p diverge. This paper aims at filling this gap.

The oracle property of the penalized likelihood based variable selection methods appears to be attractive, as it allows one to adapt to the unknown zero restrictions without paying a price. However, it parallels the super-efficiency property of the Hodges estimator, which in its simplest form is a hard-thresholding estimator exhibiting sparsity and the oracle property. Leeb and Potscher (2005) showed that the oracle property is an asymptotic feature that holds only point-wise in the parameter space, and the estimators that have the oracle properties have poor properties in minimax mean squared error and construction of rate-optimal confidence intervals. Furthermore, estimators possessing the oracle property are not exempt from the Hajek-LeCam local asymptotic minimax theorem. Most of these results usually deal with the classical fixed p situation. It remains an interesting

question as to what happens in the divergent p scenario.

A natural question that arises is whether valid hypothesis testing can be performed based on these penalized likelihood procedures that yield sparse estimators and have the oracle variable selection property. Hypothesis testing often proceeds by constructing a test statistic based on a consistent estimator, which in turn yields a consistent testing procedure. In this context, there has been considerable recent interest in testing against sparse alternatives. See for example Ingster (1998, 1997); Ingster and Suslina (2003); Donoho and Jin (2004); Hall and Jin (2010); Cai et al. (2011); Arias-Castro et al. (2011); Ingster et al. (2010) for details. One natural question is whether one can construct consistent tests against sparse alternatives by using sparse estimators that are consistent for variable selection. Here we are interested in the problem of high dimensional sparse alternatives. That is we not only have $p \to \infty$, but also have a priori knowledge that we are testing against sparse alternatives.

In the context of testing against sparse alternatives, numerous authors have studied the sharp detection boundary (Ingster, 1998, 1997; Ingster and Suslina, 2003; Donoho and Jin, 2004; Hall and Jin, 2010; Arias-Castro et al., 2011; Ingster et al., 2010). In particular, Arias-Castro et al. (2011) and Ingster et al. (2010) consider the global hypothesis testing problem in Gaussian linear models. They provide the detection boundary of the problem under different regimes of sparsity and conditions on the design matrix $\boldsymbol{X}$. Arias-Castro et al. (2011) also analyze the performance of the usual likelihood ratio test for fixed design matrices that satisfy some low coherence conditions.

In this paper, we study both the usual likelihood ratio test and the tests using penalized likelihood estimators that perform consistent variable selection, for random sub-Gaussian design matrices. We compare their asymptotic properties under different regimes of sparsity. Intuitively, since variable selection is a more difficult problem than hypothesis testing, one pays a price and can possibly lose power depending on the degree of sparsity. In this context, we derive the asymptotic order of detection boundaries for the tests based on concave penalized oracle variable selection estimators under different regimes of sparsity. Here by detection boundary of a test we mean the necessary alternative signal strength for asymptotic consistency of any test.

We show that the usual non-penalized least squares based test or the usual likelihood ratio (LR) test performs well in the dense regime (defined in Section 2), whereas tests based on many popular penalized estimators which perform consistent variable selection falls well short in terms of testing errors. Similar suboptimal performance of these penalized likelihood based tests compared to the LR test is also observed under local alternatives. In this context, we also characterize local alternatives in the context of sparse alternatives with a diverging number of parameters. In the case of the sparse regime (defined in Section 2), however, the roles of these two are reversed. Specifically, the LR test has suboptimal performance but the penalized consistent variable selection estimator based tests perform much better.

The paper is arranged as follows. In Section 2, we introduce and formulate the problem. In Section 3, we introduce several classes of tests, including the Oracle test, the LR test, and the penalized likelihood based tests. In Section 4, we study the size of the tests discussed in Section 3. In Section 5, we study the detection boundary of the aforementioned tests, analyze the oracle approximation property of the penalized likelihood based tests and also study the power of the tests against local alternatives. In Section 6, we discuss the validity of the results using simulations. We collect all the proofs into Appendix A.

1.2 The Testing Problem

Consider the regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y} = (y_1 \dots y_n)^T \in \mathbb{R}^n$ is a vector of n observed outcomes, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ is an unknown parameter of interest, and $\epsilon \in R^n$ is a vector of independent and identically distributed Gaussian variables with mean 0 and variance σ^2. Throughout this paper, we assume that σ is known and hence assume it to be 1 without loss of generality. Here we consider the scenario that $\boldsymbol{X}$ is a random design matrix, i.e. $(y_i, \boldsymbol{x}_i^T)$ are independent and identically distributed, where $\boldsymbol{x}_i$ is a $p \times 1$ vector and ϵ_i is independent of $\boldsymbol{x}_i$. We denote the probability law of the data parametrized by $\boldsymbol{\beta}$ by $\mathbb{P}_{\boldsymbol{\beta}}$. We allow $p \to \infty$ and calibrate our asymptotics as $n := n(p) \to \infty$. Let

$M(\boldsymbol{\beta}) = \sum_{j=1}^{p} I(\beta_j \neq 0)$ denote the number of non-zero coefficients in $\boldsymbol{\beta}$ and $R_k^p = \{\boldsymbol{\beta} \in \mathbb{R}^p : M(\boldsymbol{\beta}) = k\}$ denote the space of $\boldsymbol{\beta}$ with the number of non-zero coefficients equal to k. We also allow $k := k(p) \to \infty$.

In this paper, we consider hypothesis testing against sparse alternatives as described below. Let $\Theta_k^A = \{\boldsymbol{\beta} \in \bigcup_{k' \geq k} R_{k'}^p : \min\{|\beta_j| : \beta_j \neq 0\} \geq A\}$ for some $A > 0$. We consider the following hypotheses:

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \text{ vs } H_{k,A} : \boldsymbol{\beta} \in \Theta_k^A.$$

We note that this types of alternatives has been considered by Arias-Castro et al. (2011), referred to as the "*Sparse Fixed Effects Model*" or SFEM. For $p \to \infty$ and $n \to \infty$, let $k = p^{1-\theta}$ where $\theta \in [0, 1)$ is called the sparsity index (Donoho and Jin, 2004). Following the convention in Cai et al. (2011), we call the regime corresponding to $\theta \leq \frac{1}{2}$ as the *Dense Regime*, which assumes moderate sparsity, and that corresponding to $\theta > \frac{1}{2}$ as the *Sparse Regime*, which corresponds to strong sparsity.

For any test ϕ , let $\alpha(\phi) = \mathbb{E}_0(\phi)$ denote the type I error and $\eta(\phi, \boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}}(1 - \phi)$ denote the type II error. Let $\eta(\phi, \Theta_k^A) = \sup_{\boldsymbol{\beta} \in \Theta_k^A} \eta(\phi, \boldsymbol{\beta})$, and $\gamma(\phi) = \gamma(\phi, \Theta_k^A) = \alpha(\phi) + \eta(\phi, \Theta_k^A)$, $\kappa(\alpha) = \inf_{\alpha(\phi) \leq \alpha} \eta(\phi, \Theta_k^A)$. Then $0 \leq \kappa(\alpha) \leq 1 - \alpha$. Let $\gamma = \gamma_p(A) = \inf_\phi\{\gamma(\phi, \Theta_k^A)\} = \inf_{\alpha \in (0,1)}\{\alpha + \kappa(\alpha)\}$, which is the minimax total error probability. Problems on distinguishability and detectability are related to finding conditions on $A = A_{p,n,k}$ which separate the cases $\gamma_p(A) \to 1$ (indistinguishability) and $\gamma_p(A) \to 0$ (distinguishability). In particular, the detection boundary refers to the rate of the quantity $A = A_{p,n,k}$ below which all tests are asymptotically powerless, i.e., $\gamma_p(A) \to 1$, and above which one can find an asymptotically powerful test rendering $\gamma_p(A) \to 0$. In cases where one can characterize the exact constants apart from the rates of A, the detection boundary often refers to the constant deciding the phase transition between distinguishability and indistinguishability.

Now note that, this is a subproblem of testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$. It is well known that, when $p \to \infty$ we can encounter loss of power using the standard F-test. See for example, Bai and Saranadasa (1996), Chen and Qin (2010) and Lopes and Wainwright (2011) for more details. However, the complexity of the problem changes if one assumes

that the alternative is sparse. In later sections, we will see how the powers of tests based on non-convex penalized likelihood procedures that yield sparse estimators behave depending on various combinations of (n, p, k, A), where A is used to define Θ_k^A earlier. In particular, we will see different power properties between the dense ($\theta \leq \frac{1}{2}$) and sparse ($\theta > \frac{1}{2}$) regimes.

1.2.1 Notations, Definitions and Assumptions

We provide a brief summary of the notations used throughout the paper. For two real numbers a and b, denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Similarly for a set of real numbers indexed by $\mathcal{T}$, $\{a_j : j \in \mathcal{T}\}$, denote $\wedge_{j\in\mathcal{T}} a_j = \min_{j\in\mathcal{T}}\{a_j\}$ and $\vee_{j\in\mathcal{T}} a_j = \max_{j\in\mathcal{T}}\{a_j\}$. Also denote $A = \min\{|\beta_j| : \beta_j \neq 0\}$. For an $n \times p$ matrix $\mathbb{G}$ with column vectors $\mathbf{g}_1, \ldots, \mathbf{g}_p$ and a subset $\mathcal{T} \subseteq \{1, \ldots, p\}$, denote by $\mathbb{G}_\mathcal{T}$ the matrix with column vectors $\mathbf{g}_j (j \in \mathcal{T})$. Similarly for a $p \times 1$ vector $\boldsymbol{z}$ and a subset $\mathcal{T} \subseteq \{1, \ldots, p\}$, denote by $z_\mathcal{T}$ the vector $(z_j : j \in \mathcal{T})$. For a square matrix $\mathbb{G}$, denote by $s_{\min}(\mathbb{G})$ and $s_{\max}(\mathbb{G})$ the smallest and largest eigenvalues of $\mathbb{G}$ respectively. For any set S, we use S^c to represent the complement of S. Also if a_n and b_n are two sequences of real numbers then $a_n \gg b_n$ (and $a_n \ll b_n$) implies that $a_n/b_n \to \infty$ (and $a_n/b_n \to 0$) as $n \to \infty$, respectively. Similarly $a_n \gtrsim b_n$ (and $a_n \lesssim b_n$) implies that $\liminf a_n/b_n = C$ for some $C \in (0, \infty]$ (and $\limsup a_n/b_n = C$ for some $C \in [0, \infty)$). Alternatively, $a_n = o(b_n)$ will also imply $a_n \ll b_n$ and $a_n = O(b_n)$ will imply that $\limsup a_n/b_n = C$ for some $C \in [0, \infty)$).
Following Arias-Castro et al. (2011), we say that a matrix $\mathbf{C}_{p\times p} \in S_p(\gamma, \Delta)$ if the following two conditions hold

(i) $|c_{jk}| < 1 - (log(p))^{-1}$ for every $j \neq k$

(ii) For all j, $|\{k : |c_{jk}| > \gamma\}| \leq \Delta$.

We say that a random variable X has a sub-Gaussian distribution if it satisfies any of the following three equivalent conditions. There exists constants H_i, $i = 1, 2, 3$ such that

1. $\mathbb{P}(|X| > t) \leq e^{1-t/H_1^2}$ for all $t > 0$,

2. $(\mathbb{E}(|X|^r))^{1/r} \leq H_2\sqrt{r}$ for all $r \geq 1$,

3. $\mathbb{E}(\exp(X^2/H_3^2)) \leq e$.

If X has a sub-Gaussian distribution, we will define its sub-Gaussian norm as $\|X\|_{\psi^2} := \sup_{r\geq 1}\{\mathbb{E}(|X|^r)\}^{1/r}/\sqrt{r}$, where the subscript ψ^2 is used to denote that the norm is a sub-Gaussian or ψ_2 Orlicz norm and not the usual Euclidean norm. It can be shown that up to multiplicative absolute constants $\|X\|_{\psi^2}$ is the smallest value of $H_i,\ i = 1, 2, 3$ satisfying inequalities 1, 2 and 3 in the definition of sub-Gaussian random variables above. We say that a random vector $\boldsymbol{Z} \in \mathbb{R}^k$ is sub-Gaussian if the one-dimensional marginals $\langle \boldsymbol{Z}, x\rangle$, are sub-Gaussian random variables for all $x \in \mathbb{R}^k$, where for two vectors w_1, w_2 in $\mathbb{R}^k$, $\langle w_1, w_2\rangle$ denotes the usual Euclidean inner product between them. The sub-Gaussian norm of $\boldsymbol{Z}$ is defined as $\|\boldsymbol{Z}\|_{\psi^2} := \sup_{\|x\|=1} \|\langle \boldsymbol{Z}, x\rangle\|_{\psi^2}$. Throughout the paper we say that a random matrix $\boldsymbol{X}$ is sub-Gaussian with parameters $(\boldsymbol{\Sigma}, H)$ if the $\mathbb{E}(X_{ij}) = 0, \mathbb{E}(X_{ij}^2) = 1$ for all i, j and the rows of $\boldsymbol{X}$ are i.i.d sub-Gaussian random vectors with covariance matrix $\boldsymbol{\Sigma}$ and sub-Gaussian norm of each row at most H, where $\boldsymbol{\Sigma}$ is a non-negative definite covariance matrix. For example, $X \sim N_p(0, \Sigma)$ implies $H = s_{\max}(\Sigma)$. Throughout we will also assume that H is bounded away from 0 unless specified otherwise. We will also assume that Σ is nonsingular and hence this will typically imply that if $p \leq n$, the rank of the sample covariance matrix $\boldsymbol{X}^T\boldsymbol{X}/n$ is p with probability 1.

1.3 Classes of Tests

1.3.1 The LR Test and Oracle Test

Consider the usual F-test, which is the chi-square test given a known variance, as follows. In particular, we will the call test obtained by rejecting for large values of $\|\boldsymbol{\Pi}\mathbf{y}\|_2^2$ with $\boldsymbol{\Pi}$ denoting the projection onto the column space of $\boldsymbol{X}$ as the "LR Test" since it is similar to the generalized likelihood ratio test for $p \leq n$. Since conditional on $\boldsymbol{X}$, $\|\boldsymbol{\Pi}\mathbf{y}\|_2^2 \sim \chi^2_{p\wedge n}(\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2)$, we can perform the LR test at a level α by rejecting the null hypothesis when $\|\boldsymbol{\Pi}\mathbf{y}\|_2^2 > \chi^2_{p\wedge n,1-\alpha}$. Note that when $p \leq n$, the LR test is the same as rejecting when $\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 > \chi^2_{p,1-\alpha}$ where $\boldsymbol{\beta}$ is the usual least squares estimator i.e. $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}(2n)^{-1}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$. As noted by Arias-Castro et al. (2011), for a fixed design $\boldsymbol{X}$, it is easy to show that the LR test is powerless when $\|\boldsymbol{X}\boldsymbol{\beta}\|^2/\sqrt{p\wedge n} \to 0$, and is asymptotically powerful when

this quantity tends to ∞. When $\|\boldsymbol{X}\boldsymbol{\beta}\|^2/\sqrt{p \wedge n} \to c$ for some $0 < c < \infty$, power of the LR test is strictly inside $(0, 1)$. Hence the power of the LR test depends on the order of $\|\boldsymbol{\beta}\|^2$. To be more precise, we shall see that, the signal strength that is most difficult for the LR test to detect is given by $\|\boldsymbol{\beta}\|^2 = kA^2$. In the nearly orthogonal design if $kA^2/\sqrt{p \wedge n}$ diverges to ∞, we need A to grow at a faster rate in the sparse regime than the dense regime.

On the other hand, suppose now we know the location of the signals O. Then we can ignore the columns of $\boldsymbol{X}$ that corresponds to O^c, the compliment of the set O, and perform a test by rejecting the null hypothesis for large values of $\|\mathbf{\Pi_O y}\|_2^2$ with $\mathbf{\Pi_O}$ denoting the projection onto the column space of $\boldsymbol{X}_O$. For $k \leq n$, this is equivalent to rejecting when $\|\boldsymbol{X}_O\widehat{\boldsymbol{\beta}}_O\|^2 > \chi^2_{k,1-\alpha}$, where $\widehat{\boldsymbol{\beta}}_O = \text{argmin}_{\boldsymbol{\beta}_O} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}_O\boldsymbol{\beta}_O\|^2$. We will call this test the "Oracle test". The Oracle test is introduced as a benchmark since it is the optimal test in a minimax sense. The power of the Oracle test also depends on $\|\boldsymbol{\beta}\|^2$, but has a lower degrees of freedom since $k < p$. Of course, we do not know O and hence the the Oracle test will only serve as a benchmark against each fixed sparse signal in the alternative.

1.3.2 Tests Based On Sparse Non-Convex Penalized Likelihood Estimators

Consider the following penalized likelihood

$$Q(\boldsymbol{\beta}; p_\lambda) = \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2}{2n} + \sum_{j=1}^p p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot) : (-\infty, \infty) \to [0, \infty]$ is a penalty function and λ is a regularization parameter. For proper choices of the concave penalty function $p_\lambda(\cdot)$, such as MCP and SCAD, and the tuning parameter λ, under certain sparse eigenvalue type conditions on the design matrix, the minimizer of $Q(\boldsymbol{\beta}; p_\lambda)$ recovers the sparsity pattern. A natural question is whether one can construct tests based on these sparse estimators for testing against sparse alternatives. In particular, suppose $\widehat{\boldsymbol{\beta}}(p_\lambda)$ is the penalized likelihood estimator corresponding to p_λ, i.e., $\widehat{\boldsymbol{\beta}}(p_\lambda)$ is the minimizer of $Q(\boldsymbol{\beta}; p_\lambda)$. We will write $\widehat{\boldsymbol{\beta}}(p_\lambda)$ as $\widehat{\boldsymbol{\beta}}_\lambda$ when there is no confusion. We are interested in studying the properties of tests for H_0 in (1.2) based on $\widehat{\boldsymbol{\beta}}_\lambda$. We define this more precisely below.

Definition 1.1. *A test statistic based on $\widehat{\boldsymbol{\beta}}_\lambda$ is a measurable function $f(y, X, \widehat{\boldsymbol{\beta}}_\lambda)$ which satisfies $f(y, X, \boldsymbol{\beta}) = 0$ iff $\boldsymbol{\beta} = 0$.*

Hence, by tests based on $\widehat{\boldsymbol{\beta}}_\lambda$ we will refer to tests which reject a null hypothesis when $f(y, X, \widehat{\boldsymbol{\beta}}_\lambda) \in W$ for some Borel set $W \in \mathbb{R}$ where one has by Definition 1.1 that $0 \notin W$. Equivalently, by measurability of f, tests based on $\widehat{\boldsymbol{\beta}}_\lambda$ will be identified with rejecting a null hypothesis when $\widehat{\boldsymbol{\beta}}_\lambda \in C$ where $C \subseteq \mathbb{R}^p$ such that $\mathbf{0} \notin C$ almost surely. When there is no confusion, we call any such C a rejection region for tests based on $\widehat{\boldsymbol{\beta}}_\lambda$. A more detailed discussion on the choice of local minimizer $\hat{\boldsymbol{\beta}}_\lambda$ in the case of existence of multiple local minimizer can be found in Appendix A. However, the above definition immediately implies the following bound on the rejection probability of the tests based on $\widehat{\boldsymbol{\beta}}_\lambda$.

Proposition 1.1. *Let $C \subseteq \mathbb{R}^p$, such that $\mathbf{0} \notin C$ almost surely, be any rejection region for a test based on $\widehat{\boldsymbol{\beta}}_\lambda$. Then the rejection probability is bounded by $\mathbb{P}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}_\lambda \neq \mathbf{0})$.*

Owing to Proposition 1.1, since for any rejection region $\mathbf{0} \notin C \subset \mathbb{R}^p$ one has $\mathbb{P}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_\lambda \in C) \leq \mathbb{P}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}_\lambda \neq \mathbf{0})$, we will try to upper bound $\mathbb{P}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}_\lambda \neq \mathbf{0})$ whenever we need a upper bound on the type I or the power of the tests based on $\widehat{\boldsymbol{\beta}}_\lambda$.

Now we provide some examples of such penalized likelihood based tests. A class of tests that is similar in essence to the Oracle test and the LR test is the corresponding penalized likelihood ratio test defined as $T^{NPLRT}(p_\lambda) := sup\{Q(\boldsymbol{\beta}; p_\lambda)\} - Q(0; p_\lambda)$. We note that, $T^{NPLRT}(p_\lambda)$ satisfies Definition 1.1 as a valid test statistic if $p \leq n$. Other examples of test statistics satisfying Definition 1.1 include the quadratic statistic $T^{QNPL}(p_\lambda) := \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda\|^2$ when $p \leq n$ and $\sum_{j=1}^p \mathbf{I}(\hat{\beta_{\lambda,j}} \neq 0)$, $\|\hat{\boldsymbol{\beta}}_\lambda\|_2^2$ etc. For the sake of brevity, we will call tests based on $\widehat{\boldsymbol{\beta}}_\lambda$ as Non-convex Penalized Likelihood tests or NPL tests, and denote them by T^{NPL}. We will see in the subsequent sections that if one constructs a test based on consistent variable selection procedures, then the lower bound on signal strength required to perform variable selection also serves as a lower bound necessary to do consistent global testing. In order to perform testing, we however need to find or bound the size of this class of tests. We provide general results regarding upper bounds on size of these tests in Section 1.4.

1.3.3 Penalty Functions and Choice of Tuning Parameter

We now state the class of penalty functions to be allowed in our study. We will need the following condition on the penalty functions throughout this paper.

(C1) *The function $\rho(t)$, where $\rho(t;\lambda) = \lambda^{-1}p_\lambda(t)$ (we denote $\rho(t;\lambda) = \rho(t)$ when there is no confusion), is increasing and concave in $t \geq 0$ and has a continuous derivative $\rho'(t)$ for $t > 0$ with $\rho'(0+) > 0$. If $\rho(t)$ depends on λ, $\rho'(t;\lambda)$ is increasing in $\lambda > 0$ and $\rho'(0+)$ is independent of λ.*

Fan and Li (2001) advocated penalty functions that give estimators with three desired properties, namely, unbiasedness, sparsity and continuity. Condition **(C1)** is related to these properties. Fan and Li (2001) argue that at least in the case of an orthogonal design matrices, the following are true: (1) unbiasedness requires that the derivative $p_\lambda^{'}(t)$ is close to zero when $t \in [0,\infty)$ is large, (2) sparsity requires $p_\lambda^{'}(0+) > 0$, and (3) continuity with respect to data requires that the function $t + p_\lambda^{'}(t)$, $t \in [0,\infty)$ attains its minimum at $t = 0$. The concavity of ρ in Condition **(C1)** entails that $\rho^{'}(t)$ is decreasing in $t \in [0,\infty)$. Thus penalties satisfying Condition **(C1)** and $\lim_{t\to\infty}\rho^{'}(t) = 0$ enjoy unbiasedness and sparsity. However, the continuity does not generally hold for all penalties in this class. The SCAD penalty given by $p_\lambda^{'}(t) = \lambda\{I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda}I(t > \lambda)\}$ for some $a > 2$ and the MCP penalty given by $p_\lambda^{'}(t) = \frac{(a\lambda - t)_+}{a}$ with $a \geq 1$ satisfy Condition **(C1)** and the above three properties simultaneously. Although the L_1 penalty satisfies Condition **(C1)** as well as sparsity and continuity, it does not enjoy the unbiasedness property, since its derivative is identically equal to one regardless of $t \in [0,\infty)$. However, all of our results with the exception of Section 5.2, goes through for any penalty satisfying **(C1)** and hence in particular for the Lasso penalty.

Most of the results in the following sections are based on the NPL tests when the NPL procedures are tuned according to the universal threshold $\lambda = \sqrt{2log(p)/n}$ (Donoho and Johnstone, 1994). One of the perspectives on selecting the tuning parameter comes from the philosophy that no variable selection procedure should select extra variables than the truth. Continuing on this philosophy, one can desire that when the true underlying parameter vector is identically zero, then a reasonable procedure should not select any variable with high probability. In recent literature(Zhang and Zhang, 2012; Pan and Zhang,

2013), a similar but stronger condition has been referred to as the "null-consistency" condition. Translating to our context, this demands that under H_0, the NPL procedures do not select any variable with high probability. In particular, we have the following definition.

Definition 1.2. *λ satisfies null-consistency condition, if* $\mathbb{P}_{\mathbf{0}}(\hat{\boldsymbol{\beta}}_\lambda \neq \mathbf{0}) \to 0$.

The following proposition provides necessary rates on λ for satisfying the null-consistency condition.

Proposition 1.2. *Assume condition* **(C1)** *holds. Suppose $\boldsymbol{X}$ is sub-Gaussian with parameters $(\boldsymbol{\Sigma}, H)$ such that $\boldsymbol{\Sigma} \in S_p(\gamma, 1)$ with $\gamma = O(\frac{1}{(log(p))^{2+\epsilon}})$ for some $\epsilon > 0$. Also assume that $n \gg \max(H^4 log(p), (log(p))^{5+2\epsilon})$. If $\limsup_{p\to\infty} \frac{\lambda}{\sqrt{\frac{2log(p)}{n}}} < \frac{1}{\rho^{'}(0+)}$, then λ does not satisfy null-consistency condition.*

As shown in Proposition 1.2, the necessary rate of tuning parameter for null-consistency is often lower bounded by the universal tuning parameter $\sqrt{\frac{2log(p)}{(\rho^{'}(0+))^2 n}}$. Since for Lasso, SCAD and MCP penalty we have $\rho^{'}(0+) = 1$, the proposition above shows that, if the true parameter vector is identically zero, then for $\limsup_{p\to\infty} \frac{\lambda}{\sqrt{\frac{2log(p)}{n}}} < 1$ with probability going to 1, the NPL procedures with Lasso, SCAD or MCP penalty, selects more than the zero vector with high probability. Hence in the rest of the paper, to ensure null consistency, our choice of λ will be at least of the magnitude of $\sqrt{\frac{2log(p)}{n}}$ and most of the time we will work with $\lambda \geq \frac{1}{\rho^{'}(0+)}\sqrt{\frac{2(1+\epsilon_p)log(p)}{n}}$ for some sequence $\epsilon_p > 0$. This condition on the NPL tests will distinguish them from the LR test which can be typically thought of as the case when $\lambda = 0$. In particular, in our general discussion about NPL tests we will mostly assume tuning by $\lambda \gtrsim \frac{1}{\rho^{'}(0+)}\sqrt{\frac{2(1+\epsilon_p)log(p)}{n}}$ for some sequence $\epsilon_p > 0$ unless mentioned otherwise. It is worth noting that the existing literature in variable selection using non-convex penalized procedures imposes similar or stronger assumptions on the tuning parameter.

1.4 Size of the Tests

One of the main properties of the NPL tests is that one cannot achieve an exact asymptotic level α for $\alpha > 0$ if the penalties satisfy property **(C1)** stated earlier and the tuning parameter is selected as discussed earlier. In particular, we show that the size of the NPL tests is asymptotically 0. Indeed, this can be immediately seen if one imposes the null-consistency condition in Definition 1.2. However, under sub-Gaussian design matrices, we now give an upper bound on the type I error of NPL tests which allows us to understand the rate at which the size of the NPL tests converge to zero. This is unlike the LR test and the Oracle test where one can construct an exact level $\alpha > 0$ test. In general, in order to perform a test of a hypothesis, we need to define a critical or rejection region in a way that it controls the type error at a reasonable level. Classically, one controls the type I error of a test at a positive exact level α or at an asymptotic level α, and then compares the powers of the tests having the same asymptotic size. In general, one studies consistency of level α tests and thereafter compares asymptotic relative efficiencies of consistent tests. The following Theorem 1.1 states the property of the size of the NPL tests discussed in Section 1.3.2 .

Theorem 1.1. *(a) Under Condition (C1), for an arbitrary random design matrix $\boldsymbol{X}$, for any set $\mathbf{0} \notin C_p \subset \mathbb{R}^p$ a.s.,*

$$\mathbb{P}_{\mathbf{0}}\{\widehat{\boldsymbol{\beta}}_\lambda \in C_p\} \leq 2p\{\exp\,(-\frac{n^2\lambda^2(\rho'(0+;\lambda))^2}{2D_n}) + \mathbb{P}(B_n^c)\},$$

where $B_n = \{\max_{1\leq j\leq p} \|\boldsymbol{x}_j\|^2 \leq D_n\}$ with $\boldsymbol{x}_j$ being the jth column of $\boldsymbol{X}$ and D_n is any sequence of positive real numbers.

(b) Suppose $\boldsymbol{X}$ is sub-Gaussian with parameters (Σ, H). If p_λ satisfies Condition (C1), then for any set $\mathbf{0} \notin C_p \in \mathbb{R}^p$ a.s and all $1 > \epsilon > 0$,

$$\mathbb{P}_{\mathbf{0}}\{\widehat{\boldsymbol{\beta}}_\lambda \in C_p\} \leq 2p\{\exp\,(-\frac{n\lambda^2(\rho'(0+;\lambda))^2}{2(1+\epsilon)}) + \exp^{(-\frac{M\epsilon^2}{H^4}n)}\},$$

where $M > 0$ is a constant.

Corollary 1.1. *Suppose $n \gg H^4 log(p)$ and $\lambda \geq \frac{1}{\rho'(0+)}\sqrt{\frac{2(1+\epsilon_p)log(p)}{n}}$ for some sequence $\epsilon_p > 0$. Then under assumptions of Theorem 1.1(b), there exists $\epsilon_p \to 0$ slowly enough such that λ satisfies null-consistency condition and hence the size of any NPL test is asymptotically 0.*

Remark 1.1. Theorem 1.1(b) and Corollary 1.1 says that for sub-Gaussian design matrices, if we want to construct a test based on $\widehat{\boldsymbol{\beta}}_\lambda$, then for any rejection region C_p such that $\mathbf{0} \notin C_p$ almost surely, the size of the test is asymptotically zero when $n \gg H^2 log(p)$. The choice of $\mathbf{0} \notin C_p$ is made since one expects such a scenario to happen only under the null.

Remark 1.2. The results in Theorem 1.1(a) are quite general in the sense that it does not assume anything about the design matrix $\boldsymbol{X}$. Hence there are no restrictions on the set B_n and the real numbers D_n. Provided one can find suitable D_n for which the quantity on the right hand side of the inequality is small enough, one will get a tighter bound. This is exactly what we obtain when the rows of $\boldsymbol{X}$ are iid sub-Gaussian random vectors in Theorem 1.1(b). In this case, the suitable D_n turns out to be $O(n)$, since when the rows of $\boldsymbol{X}$ are from an i.i.d sub-Gaussian distribution then the maximum of the column norms behaves like $\sqrt{n}$ with a very high probability. This is also similar to the conditions in Fan and Lv (2011), where they assume that the columns have norm equal to $\sqrt{n}$.

Remark 1.3. Theorem 1.1 is a finite sample result and assumes nothing on the tuple (n, p, k). As in Corollary 1.1, one can of course take limsup on both sides of the inequalities to obtain large sample results.

The results in Theorem 1.1 suggest that the only way to achieve an asymptotic level $\alpha > 0$ using penalized likelihood based test is by adding an independent Bernoulli(α) to the procedure, which might be an artificial solution. Hence, for example, if we want to reject for large values of T^{NPL}, one can simply choose any non-negative sequence δ_p (which can be made to depend on the data if needed) and reject when T^{NPL} exceed δ_p. The level of such a sequence of tests is asymptotically 0 when $n\lambda^2(\rho'(0+;\lambda))^2 \to \infty$ at a proper rate. It is worth noting that the tuning parameters suggested in literature (Fan and Peng, 2004; Fan and Lv, 2011; Lv and Fan, 2009; Zhang, 2010) satisfy this rate.

Since the NPL tests have asymptotic type I error 0, we need to compare the tests at the asymptotic level 0. To accomplish this, we perform the LR test as follows. We reject the null hypothesis when $\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 > \chi^2_{p,1-\alpha_p}$, where $\alpha_p = o(1)$ is some sequence. Now all tests are of asymptotic level 0, we will investigate when the total testing errors of different tests converge to 0 or 1.

Define the total error of a test $T = I\{f(y, \boldsymbol{X}) \in C\}$ by $R_{\boldsymbol{\beta}}(T) = \mathbb{P}_{H_0}(T = 1) + \mathbb{P}_{\boldsymbol{\beta}}(T = 0)$ where $f(y, \boldsymbol{X})$ is any measurable function of the data and C is any rejection region. We denote the Oracle test and the LR Test by T^{Oracle} and T^{LR}, respectively, to be consistent with notation of the NPL tests T^{NPL}. Hence for the Oracle, LR and NPL tests, the quantities $R_{\boldsymbol{\beta}}(T^{Oracle})$, $R_{\boldsymbol{\beta}}(T^{LR})$ and $R_{\boldsymbol{\beta}}(T^{NPL})$ stand for their total testing errors respectively. Denote the maximum total error of testing of any testing procedure T by

$$Risk(T) := \mathbb{P}_0(T = 1) + \max_{\boldsymbol{\beta} \in \Theta_k^A} [\mathbb{P}_{\boldsymbol{\beta}}(T = 0)].$$

We will say that a test T is *asymptotically powerful* if $Risk(T) \to 0$ and *asymptotically powerless* if $Risk(T) \to 1$. In the following sections, we will study different regimes of sparsity which classify the tests into different categories of asymptotic power.

In subsequent sections, when we say that a test is asymptotically powerful, we mean that it is possible to choose an appropriate rejection region such that the test based on the procedure where we reject the null hypothesis outside the chosen critical region, is asymptotically powerful. Moreover when we say that a test is powerless, we mean that for every chosen critical region, the test based on that critical region is powerless.

1.5 Power Properties of the LR Test and the Non-Convex Penalized Likelihood Tests

In order to study the power properties of the classes of tests introduced in Section 3, we divide our study into three subsections. In Section 5.1, we analyze the worst case testing errors of the LR test and the NPL tests, i.e., we study $\mathrm{Risk}(T^{LR})$ and $\mathrm{Risk}(T^{NPL})$. We define the detection boundary of a test by the minimum magnitude of A required for a test to be asymptotically powerful. We provide the detection boundary of the NPL tests and compare it with the detection boundary of the LR test. Our results show that the structure of the respective detection boundaries implies that the asymptotic relative performance of the LR test and the NPL test is different according to whether $\theta \leq \frac{1}{2}$ (Dense Regime) or $\theta > \frac{1}{2}$ (Sparse Regime).

1.5.1 Detection Boundary

The worst case errors of testing, $\mathrm{Risk}(T^{LR})$ and $\mathrm{Risk}(T^{NPL})$, differ substantially in the dense regime and the sparse regime. The detection boundary of the LR test for design matrices satisfying some low correlation conditions among its columns has been previously studied by Arias-Castro et al. (2011) for fixed design matrices. We use their results to accommodate random design matrices with sub-Gaussian distributed rows. The detection boundary for the NPL tests however has not been studied before and we provide it as part of Theorem 1.2.

Theorem 1.2. *Suppose $\boldsymbol{X}$ is sub-Gaussian with parameters $(\boldsymbol{\Sigma}, H)$and assume that the NPL tests are tuned by $\lambda = \frac{1}{\rho'(0+)}\sqrt{\frac{2(1+\epsilon_p)log(p)}{n}}$ for some sequence $\epsilon_p > 0$. Recall that $\theta \in [0,1)$ is the sparsity index where $k = p^{1-\theta}$.*

(a) Suppose $\boldsymbol{\Sigma} \in S_p(\gamma, 1)$ such that $\gamma \ll p^{\theta-1}$ and that $H^4 p^{1-\theta} log(p) \ll n$. Then there exists $\epsilon_p \to 0$ slow enough such that the following holds.
(i) If $A \gg \sqrt{\frac{log(p)}{n}}$ then all NPL tests are asymptotically powerful. If $A \ll \sqrt{\frac{log(p)}{n}}$ then all NPL tests are asymptotically powerless.
(ii) Suppose $\theta > \frac{1}{2}$ and let $A = \sqrt{\frac{2tlog(p)}{n}}$. If $t < (1 - \sqrt{(1-\theta)})^2$ then all NPL tests are asymptotically powerless. If $t > (1 - \sqrt{(1-\theta)})^2$ then all NPL tests are asymptotically powerful.

(b) Suppose $\boldsymbol{\Sigma} \in S_p(\gamma, 1)$ such that $\gamma \ll (log(p))^{-1}$ and that $n \gg H^4 (log(p))^3$. If $A \gg \frac{(p \wedge n)^{1/4}}{\sqrt{kn}}$ then the LR test is asymptotically powerful. If $A \ll \frac{(p \wedge n)^{1/4}}{\sqrt{kn}}$ then the LR test is asymptotically powerless.

Remark 1.4. *Theorem 1.2 has interesting implications about the relation between signal strengths required for successful detection by the NPL tests and the LR test respectively. In particular, we note that under conditions of Theorem 1.2, whenever $p \leq n$ and $H \ll p^{\delta}$ for all $\delta > 0$, one has $\sqrt{log(p)/n} \gg (p \wedge n)^{1/4}/\sqrt{kn}$ when $\theta \leq \frac{1}{2}$, i.e. the dense regime, and $\sqrt{log(p)/n} \ll (p \wedge n)^{1/4}/\sqrt{kn}$ when $\theta > \frac{1}{2}$, i.e. the sparse regime. Hence Theorem 1.2 implies that the performance of the NPL tests and the LR test is reversed between the dense and the sparse regimes, at least when $p \leq n$. The requirement of $p \leq n$ is for technical reasons and is partially due to control over random design matrices. In particular, for $p \leq n$,the NPL tests require*

Overview			
Regime	Signal Strength A	LR Test	NPL Tests
Dense($\theta \leq \frac{1}{2}$)	$A \gg \sqrt{\frac{log(p)}{n}}$	Powerful	Powerful
	$\frac{(p\wedge n)^{1/4}}{\sqrt{kn}} \lesssim A \ll \sqrt{\frac{log(p)}{n}}$	Powerful	Powerless
	$A \ll \frac{(p\wedge n)^{1/4}}{\sqrt{kn}}$	Powerless	Powerless
Sparse($\theta > \frac{1}{2}$)	$A \gg \frac{(p\wedge n)^{1/4}}{\sqrt{kn}}$	Powerful	Powerful
	$\sqrt{\frac{log(p)}{n}} \lesssim A \ll \frac{(p\wedge n)^{1/4}}{\sqrt{kn}}$	Powerless	Powerful
	$A \ll \sqrt{\frac{log(p)}{n}}$	Powerless	Powerless

Table 1.1: Summary of the performance of the LR Test and the NPL Tests tuned by $\lambda = \sqrt{\frac{2(1+\epsilon_p)log(p)}{n}}$ ($\epsilon_p > 0$ converging to 0 at a sufficiently slow rate) for different sparsity regimes and classes of alternatives, where A stands for the minimum signal strength of the nonzero signals.

more signal strength to be asymptotically powerful than the LR test in the dense regime, while the LR test requires more signal strength for being asymptotically powerful than the NPL tests in the sparse regime. The results are summarized in Table 1.

Remark 1.5. *According to (a(ii)) of Theorem 1.2 , the exact constant of the detection boundary can be evaluated in the sparse regime and equals the detection boundary of the minimum p-value test (Arias-Castro et al., 2011). We note that the Higher Criticism test (Donoho and Jin, 2004; Arias-Castro et al., 2011) is sharp optimal for all $\theta > 1/2$. Comparing the detection boundary of the NPL tests with the Higher Criticism test, (a(ii)) of Theorem 1.2 also implies that the NPL tests are sharp optimal and is equal in performance to the Higher Criticism test when $\theta \geq 3/4$ and the is suboptimal in terms of constants compared to the Higher Criticism test when $1/2 < \theta < 3/4$.*

Theorem 1.2 also suggests that one can construct an omnibus test which is adaptive over different classes of sparsity by combining the LR test and the NPL test. Since for $p \leq n$, the LR test is an extreme case of the NPL test by setting the tuning parameter $\lambda = 0$, this means the omnibus test combines the tests with different values of the tuning parameter λ. Specifically, consider any estimator $\hat{\boldsymbol{\beta}}_\lambda$ with the penalty satisfying condition **(C1)**. Further assume that $p_\lambda(t) = 0$ when $\lambda = 0$, which is typically satisfied by many commonly used penalties for variable selection in high-dimensional regression, such as

Lasso, MCP, and SCAD. Recall the quadratic NPL test statistic

$$T^{QNPL}(p_\lambda) := \|\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\|^2.$$

Note that when $\lambda = 0$, this corresponds to the LR test. The omnibus test is defined by combining $T^{QNPL}(p_\lambda)$ over suitable values of λ as follows

$$\text{Reject when : } \max\left\{\mathbf{I}\left(T^{QNPL}(p_\lambda) > t_p(\lambda)\right) : \lambda \in \left\{0, \sqrt{\frac{2(1+\epsilon)log(p)}{\rho'(0+)n}}\right\}\right\} > 0 \qquad (1.1)$$

for some deterministic sequence $t_p(\lambda)$ to be decided later. In particular, we reject the global null if either of the LR test or NPL test rejects the null. Then we have the following theorem where we will assume $p \leq n$ to ensure that the ordinary least square is well defined. Also we provide the results in terms of the Lasso penalty for simplicity. Similar result holds for a general class of concave penalties by invoking results in Zhang and Zhang (2012).

Theorem 1.3. *Suppose that the rows of $\boldsymbol{X}$ are i.i.d multivariate Gaussian with mean $\mathbf{0}$ and covariance matrix Σ with $\Sigma_{jj} = 1\ \forall j$. Assume that $\Sigma \in S_p(\gamma, 1)$ with $\gamma \ll (log(p))^{-1}$, $\frac{1}{s_{\min}(\Sigma)} = O(1)$ and $\max(s_{\max}(\Sigma), H) \ll p^\delta$ for all $\delta > 0$. Finally assume penalization by Lasso penalty i.e. $p_\lambda(t) = \lambda t$. Then there exists deterministic $\epsilon > 0$ and sequence $t_p(\lambda)$ such that the omnibus test given by equation* (1.1) *is optimally adaptive over different regimes of sparsity, i.e. it is asymptotically powerful whenever $A \gg \frac{(p\wedge n)^{1/4}}{\sqrt{kn}}$ in the dense regime and $A \gg \sqrt{\frac{log(p)}{n}}$ in the sparse regime.*

Remark 1.6. *Theorem 1.3 suggests that there exist theoretical values of λ that yield an an optimal test that is adaptive to both sparse and dense regimes. However, it is an interesting future research question if there exists data driven way of selecting λ to yield the optimal test.*

1.5.2 Approximation of the Oracle Test

The optimality criterion in the previous subsection is based on worst case risk considerations where one needs to analyze the worst case risk under various classes of alternatives, and the signal locations are allowed to vary over coordinates of $\boldsymbol{\beta}$. To study a different

point of view, recall that in Section 1.3.1 the Oracle test was introduced as a benchmark when one knows the possible locations of the signals in the alternative. It is of interest to know if there exists a testing procedure which mimics the performance of the Oracle test without knowing the locations of the signals.

In this section, we will show that the risk of the test based on $T^{NPLRT}(p_\lambda)$ and $T^{QNPL}(p_\lambda)$, introduced in Section 1.3.2, mimics the risk of the Oracle test under suitable regularity conditions. In particular, note that since $T^{NPLRT}(p_\lambda) = -\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda\|^2 + 2\boldsymbol{y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda - 2n\sum_{j=1}^p p_\lambda(|\widehat{\boldsymbol{\beta}}(p_\lambda)_j|)$, where $\widehat{\boldsymbol{\beta}}_\lambda$ is the corresponding penalized likelihood estimator, $T^{NPLRT}(p_\lambda)$ is a NPL test by Definition 1 whenever $p \leq n$. Similarly, $T^{QNPL}(p_\lambda) = \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda\|^2$ is also a NPL test according to Definition 1 whenever $p \leq n$. We will show that the risks of the tests based on $T^{NPLRT}(p_\lambda)$ and $T^{QNPL}(p_\lambda)$ mimic the risk of the Oracle test if the signal strength exceeds the detection boundary. To be consistent with the notation, we denote the total testing error of NPRT and QNPL as $R_{\boldsymbol{\beta}}(T^{NPLRT}(p_\lambda))$ and $R_{\boldsymbol{\beta}}(T^{QNPL}(p_\lambda))$ respectively. To provide cleaner results, we state the following additional condition on the penalty function used in the NPL estimator.

(C1′) The penalty $p_\lambda(t)$ satisfies condition **(C1)**, $0 < \rho'(0+;\lambda) < \infty$ and $0 < p'_\lambda(t) = 0$ for $t > c\lambda$ for some constant $c > 0$. In that case we say that $p_\lambda(t)$ satisfies condition **(C1′)** with constant c.

Examples of penalty functions satisfying **(C1′)** are SCAD and MCP penalties or any convex combination of them. In the following, we refer to penalties satisfying **(C1′)** as **(C1′)** penalties. Theorem 3 states the performance of these tests with respect to the Oracle test.

Theorem 1.4. *Suppose $\boldsymbol{X}$ is sub-Gaussian with parameters $(\boldsymbol{\Sigma}, H)$ such that $\frac{1}{s_{\min}(\boldsymbol{\Sigma})} = O(1)$ and $\max(s_{\max}(\Sigma), H) \ll p^\epsilon$ for all $\epsilon > 0$. Also suppose that the NPL tests are tuned by $\lambda = \frac{1}{\rho'(0+)}\sqrt{\frac{2(1+\epsilon_p)log(p)}{n}}$ for some sequence $\epsilon_p > 0$. Assume that condition **(C1′)** holds and that $p \leq n$. If $A \gg \sqrt{\frac{log(p)}{n}}$ and the rejection region is $C_p = (t_p, \infty)$ for some sequence t_p, the there exists $\epsilon_p \to 0$ slow enough such that*

$$R_{\boldsymbol{\beta}}(T) = (1 - \mathbb{P}_\beta(\|\boldsymbol{X}_O\widehat{\boldsymbol{\beta}}_O\|^2 > t_p)) + o(1),$$

where T is either $T^{NPLRT}(p_\lambda)$ or $T^{QNPL}(p_\lambda)$.

Remark 1.7. Theorem 1.4 suggests that, if we choose t_p to be the appropriate chi-square quantile $\chi^2_{k,1-\alpha_p}$ for any $\alpha_p \to 0$ used to construct the Oracle test, the power function of the NPLRT and QNPL tests mimic the power function of the Oracle test whenever the signal strength exceeds the detection boundary of the NPL tests, provided the tuning parameter λ is chosen to satisfy null-consistency condition and the design matrix is drawn from suitable sub-Gaussian ensembles.

Remark 1.8. *The condition $p \leq n$ is assumed for two reasons. Under $p \leq n$, the minimizer of $Q_n(\boldsymbol{\beta})$ coincides with the oracle estimator with high probability. This can be relaxed under the MCP penalty where the MC+ algorithm guarantees finding the suitable local minimum which resembles the oracle estimator with high probability. Since we work with more general penalties we omit such cases. The assumption is also used to guarantee relatively shorter arguments while controlling maximum and minimum eigenvalue of the sample covariance matrix. However, this can be relaxed since it is possible to accommodate larger p by going through restricted eigenvalue conditions.*

1.5.3 Local Alternatives

In the previous two sections, we analyzed the necessary and sufficient conditions on the signal strength for NPL and LR tests to be asymptotically powerful. Another important setting for comparing tests is to study their power against local alternatives which we now define. We will say that a sequence of probability measures $\mathbb{P}_n$ is local w.r.t another sequence of probability measures $\mathbb{Q}_n$ defined on the same probability space if the Kulback-Leibler divergence $\mathrm{KL}(\mathbb{P}_n|\mathbb{Q}_n) \in [0, 1)$, where for two probability measures $\mathbb{P}, \mathbb{Q}$ with $\mathbb{P}$ absolutely continuous with respect to $\mathbb{Q}$ one has $\limsup \mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \int log \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{P}$. By Pinsker's Inequality, this implies that the total variation distance between $\mathbb{P}_n$ and $\mathbb{Q}_n$ remains bounded between 0 and 1. Our definition of local alternatives is inspired by the fact that in classical parametric literature this is also referred to as the contiguous or local alternatives. For more details of contiguous alternatives in regular parametric theory, one might refer to Van der Vaart (2000). In this section, we compare the local power of the NPL tests and the LR test. However, for the sake of completeness, we first provide a

characterization of sparse local alternatives in the divergent p situation in the following proposition.

Proposition 1.3. *Suppose that rows of $\boldsymbol{X}$ are i.i.d from a multivariate distribution F on $\mathbb{R}^p$ such that $\mathbf{E}_F(\boldsymbol{X}^T\boldsymbol{X}/n) = \Sigma$ and let $O(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$. If $s_{\max}(\boldsymbol{\Sigma}_{O(\boldsymbol{\beta})}) \lesssim \frac{1}{n\|\boldsymbol{\beta}\|^2}$ then $\mathbb{P}_{\boldsymbol{\beta}}$ is local w.r.t $\mathbb{P}_{\mathbf{0}}$.*

In particular, if the k-sparse eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded away from ∞ then $n\|\boldsymbol{\beta}\|^2 = O(1)$ yields the local alternatives similar to regular fixed dimensional parametric case. In the following theorem, we analyze the power of the NPL tests and the LR test against alternatives satisfying $n\|\boldsymbol{\beta}\|^2 = O(1)$ and hence our results automatically covers local alternatives as characterized by Proposition 1.3 when the k-sparse eigenvalues of $\boldsymbol{\Sigma}$ are uniformly bounded away from ∞.

Theorem 1.5. *Suppose $n\|\boldsymbol{\beta}\|^2 = c$ for some constant $c > 0$. Also suppose that $\boldsymbol{X}$ is sub-Gaussian with parameters $(\boldsymbol{\Sigma}, H)$ such that $\max(s_{\max}(\Sigma), H) \ll log(p)$ and $n \gg p^{1-\theta}(log(p))^{-2}$. Finally let $\lambda \geq \sqrt{\frac{2(1+\epsilon_p)log(p)}{(\rho'(0+))^2 n}}$ for some sequence $\epsilon_p > 0$. Then there exist sequences t_p and $\epsilon_p \to 0$ slow enough such that for any set $0 \notin C_p \in \mathbb{R}^p$ a.s. one has*

$$\mathbb{P}_{\boldsymbol{\beta}}(T_n^{LR} > t_p) \gg \mathbb{P}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_\lambda \in C_p)$$

Remark 1.9. Theorem 1.5 suggests that under the local alternatives prescribed by $n\|\boldsymbol{\beta}\|^2 = c$ for some constant $c > 0$, the power of the LR Test, for an appropriate rejection region, converges to 0 at a slower rate than the power of any NPL test. The results imply that, under the assumptions of the theorem, the LR Test is locally more powerful than any NPL test.

1.6 Simulations

1.6.1 Overview

We performed simulations to compare the performance of the tests considered in this paper against alternatives of different types in both sparse and dense regimes. We divide our simulations study into dense and sparse regimes. The test statistics considered in

the simulation were the SCAD penalized likelihood ratio test, the LR test and the Oracle test respectively. For the SCAD test, the tuning parameter was selected using the BIC criterion (Dicker and Lin, 2012). Moreover in order to bypass the debate of choosing an appropriate size and critical region, we instead provide the box plots of the test statistics under the null and alternative hypotheses. If the two boxplots do not have a substantial overlap, it means one can construct a test based on the corresponding statistics to obtain asymptotic power of detection.

1.6.2 The Dense Regime

In the dense regime ($\theta \leq 1/2$), which corresponds to moderate sparsity, we considered the following sample size, parameter number and oracle size combination: $n = 500, p = 50, k = 10$. The design matrices were set with rows following iid $N(0; \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is taken to be $AR1(0.1)$. The signal strength of each coordinate was chosen to be $1/n^{\gamma/16}$, where γ was chosen differently to yield signals corresponding to (a) regions above the detection boundary for the NPL test, (b) below the detection boundary for the NPL tests but above the detection boundary of the LR test, and (c) the local alternatives. The choice of $\gamma/16$ is made so that as γ increases, the signal strength decreases from above the detection boundary of the LR test to the local alternatives.

Figure 1.1 gives the results against the alternatives corresponding to the region above the detection boundary of the NPL tests. It can be seen from the separation of the box plots under the null and alternative hypotheses that all the three test procedures (LR, NPL and Oracle tests) have consistency against such alternatives. This is consistent with Theorem 1.2.

Figure 1.2 shows the results when the alternatives are in the region below the detection boundary of the NPL tests but above the detection boundary of the LR test. The results show that, no matter how we chose the rejection region for SCAD penalized likelihood ratio test, we cannot obtain any power since the box plots under the null and alternative hypotheses completely overlap. It is also worth noting that in this case, under the null and the alternative hypothesis corresponding to the regions below the detection boundary for the NPL tests but above detection boundary of the LR test, the SCAD Test statistic puts

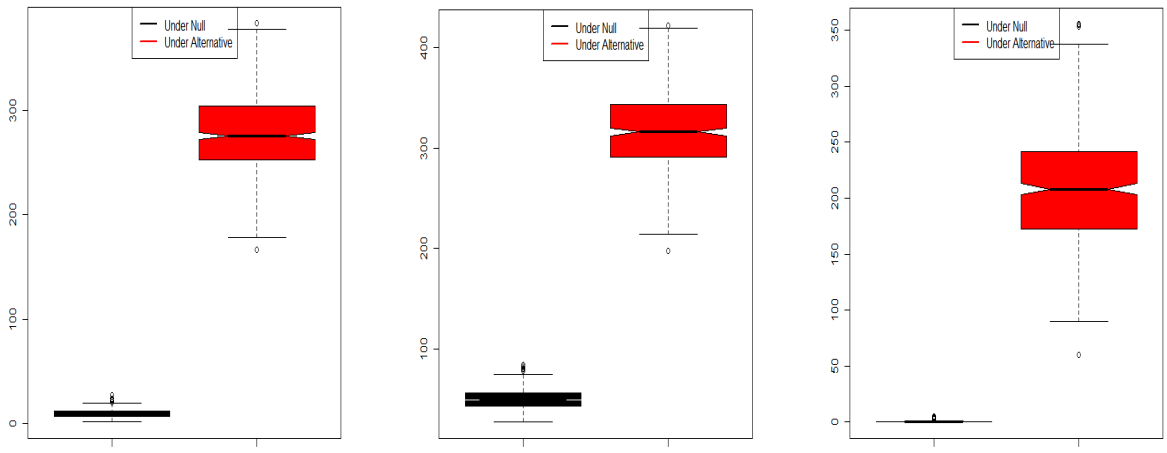

Figure 1.1: (Left To Right) The Boxplot of the Oracle test statistic, the LR test statistic and the SCAD penalized likelihood ratio test statistic over 500 simulations in the dense regime: $n = 500, p = 50, k = 10$, signal strength of each $\beta_j = n^{-\gamma/16}$, where $\gamma = 4$. The rows of $\boldsymbol{X}$ are iid $N(0, \Sigma)$ with $\Sigma = AR(0.1)$

a overwhelming mass at 0. Hence, we have added the same $Uniform(0, 1)$ noises to the statistic under both the null and alternative for a better visualization. Otherwise there will be just a straight line at 0 which is the bar plot under the null and such alternatives. Figure 1.2 also shows that for the LR test, we can choose a critical region so that we have power against these alternatives, but for the SCAD test there is no hope in this setting. This is exactly in accordance with Theorem 1.2.

Figure 1.3 gives the results corresponding to the local alternatives. For local or contiguous alternatives, Figure 1.3 shows that the overlap between the boxplots under the null and alternative hypotheses for the SCAD test is much more than the overlap for the LR Test statistic. This is consistent with Theorem 1.5.

1.6.3 The Sparse Regime

In the sparse regime $\theta > 1/2$, which corresponds to strong sparsity, we considered the following two sample size, parameter number and oracle size combinations: $n = 500, p = 100, k = 1$ for Figure 1.4 and $n = 500, p = 400, k = 1$ for Figure 1.5. The design matrices were chosen with rows following iid $N(0; I)$.

Figure 1.4 gives the results against the alternatives corresponding to the region above the detection boundary of the LR test It can be seen from the separation of the box plots

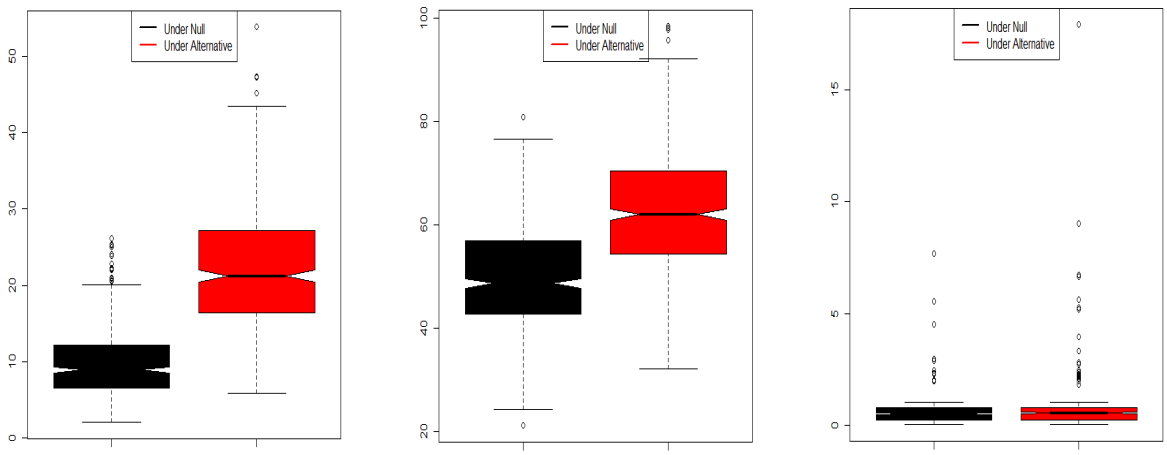

Figure 1.2: (Left To Right) The boxplot of the Oracle test statistic, the LR test statistic and the SCAD penalized likelihood ratio test statistic over 500 simulations in the dense regime : $n = 500, p = 50, k = 10$, signal strength of each $\beta_j = n^{-\gamma/16}$, where $\gamma = 8$. The rows of $\boldsymbol{X}$ are iid $N(0, \Sigma)$ with $\Sigma = AR(0.1)$

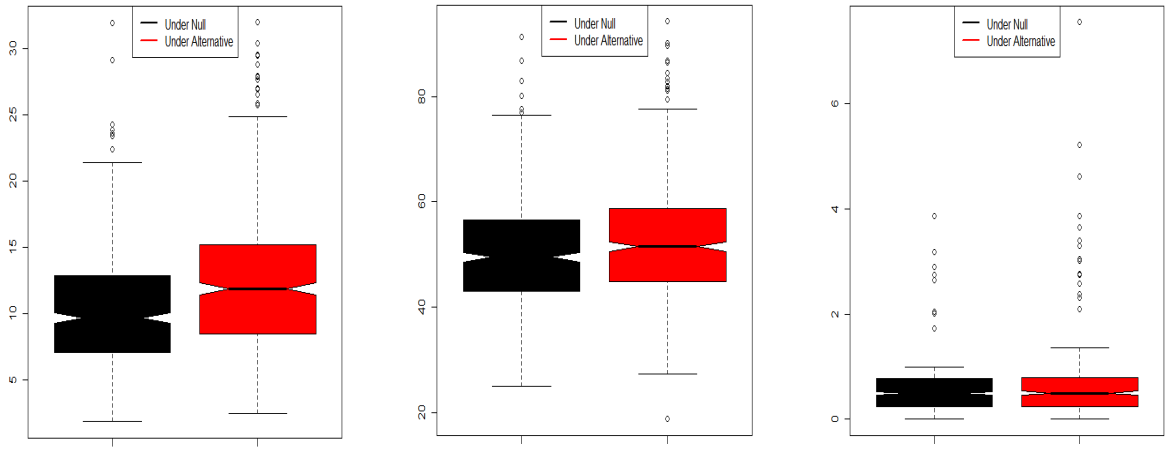

Figure 1.3: (Left To Right) The boxplot of the Oracle test statistic, the LR test statistic and the SCAD penalized likelihood ratio test statistic over 500 simulations assuming Local Alternatives : $n = 500, p = 50, k = 10$, signal strength of each $\beta_j = n^{-\gamma/16}$ where $\gamma = 10$. The rows of $\boldsymbol{X}$ are iid $N(0, \Sigma)$ with $\Sigma = AR(0.1)$

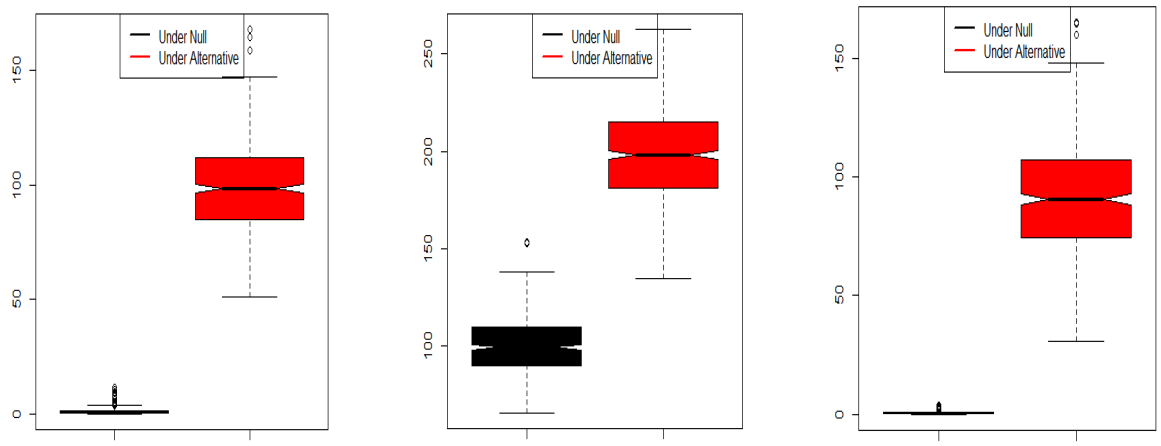

Figure 1.4: (Left To Right) boxplot of the Oracle test statistic, the LR test statistic and the SCAD penalized likelihood ratio test statistic in sparse regime 500 Simulations:$n = 500, p = 100, k = 1$, signal strength of each $\beta_j = \sqrt{\frac{p}{n}}$,rows of $\boldsymbol{X}$ are iid $N(0, I)$

that all the three test procedures (LR, NPL and Oracle tests) can be expected to have consistency against such alternatives. This is consistent with Theorem 1.2.

For simulations corresponding to Figure 1.5, the signal strength of each coordinate was chosen to be $\frac{(log(p-k))^{3/4}}{\sqrt{n}}$ to correspond to the alternatives in the region above the detection boundary of the NPL tests but below the detection boundary of the LR tests under the sparse regime. As expected from Theorem 1.2, under the sparse regime, the SCAD test has more power than the LR test which can be seen from the amount of overlap between the boxplots under the null and alternative hypotheses in Figure 1.5.

1.7 Discussions

This paper studies the asymptotic properties of the tests based on non-convex penalized likelihood (NPL) procedures which can be tuned to perform consistent variable selection and compare them with the unpenalized usual LR test under different sparsity regimes. By providing an exponential inequality, we first showed that under the null-consistency condition, the NPL tests have asymptotic size 0. We hence compare the NPL tests with the LR test that is set to have asymptotic size 0. Our results show that the performance of the LR test and the NPL tests based on consistent variable selection procedures, such as MCP and SCAD, depend on the degree of sparsity in the alternatives. The LR Test wins

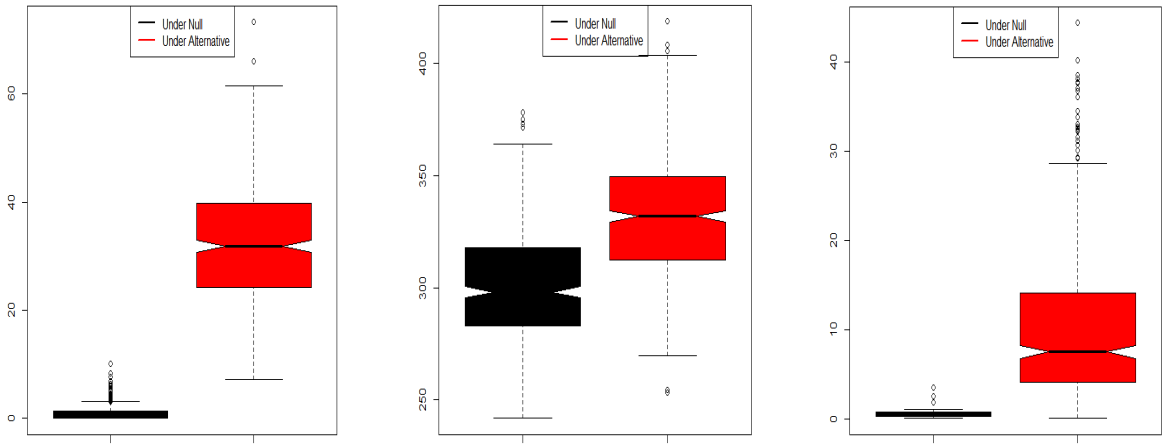

Figure 1.5: (Left To Right) boxplot of the Oracle test statistic, the LR test statistic and the SCAD penalized likelihood ratio test statistic in the sparse regime 500 Simulations:$n = 500, p = 400, k = 1$, signal strength of each $\beta_j = (log(p-k))^{3/4}/\sqrt{n}$,rows of $\boldsymbol{X}$ are iid $N(0, I)$

in the dense regime which has moderate sparsity, the NPL tests outperform the LR test in the sparse regime. These results follow from studying the detection boundaries of the LR test and the NPL tests when the penalty satisfies certain conditions. We showed that the benchmark Oracle test can be approximated by certain NPL Tests, without knowledge of the oracle set if the signal strength in the alternative is above the threshold of the detection boundary for the NPL tests. In the context of local alternatives, however, the LR test outperforms the NPL tests irrespective of the degree of sparsity.

All the results in this paper are derived under the assumption that the noise $\boldsymbol{\epsilon}$ is additive Gaussian noise. The assumption of Gaussian noise is explicitly required for determining the structure of local alternatives. Most of the results still hold under the relaxed assumption of scaled and centered sub-Gaussian noise. We note that, one can also study the properties of the tests based on Lasso and Dantzig Selector estimators using similar techniques that are used in this paper. Most of our asymptotic results are valid under the assumption of $p \leq n$ and is often due to guarantee of control over structure of random design matrices. However, we also provide some finite sample results. The conditions under which most of the results are derived are definitely not optimal and tight. At the cost of more detailed calculations, one can derive similar results under tighter conditions.

It is of future research interest to extend these results to generalized linear models. It is also of future research interest to derive the finite sample bounds if one chooses the tuning parameter by BIC instead of the theoretical value. Our simulations show that one is likely to expect similar findings regarding power to the results using the theoretically chosen tuning parameter.

Hypothesis Testing for Sparse Binary Regression

Rajarshi Mukherjee, Natesh S. Pillai* and Xihong Lin

Department of Biostatistics

Harvard School of Public Health

and

* Department of Statistics

Harvard University

2.1 Introduction

The problem of testing for the association between a set of covariates and a response has always been of fundamental statistical interest. In the context of testing for a linear relationship of covariates with a continuous response, Fisher introduced analysis of variance (ANOVA) in the 1920's, which is still widely used in the present day. In recent years, finding the detection boundary of various testing problems has gained substantial popularity. A fruitful way of finding the detection boundary is to study the minimax error of testing and obtain a threshold of signal strength under which all testing procedures in the concerned problem are useless. For Gaussian linear models, this has been extensively studied by Arias-Castro et al. (2011) and Ingster et al. (2010); these works were inspired by the previous work on hypothesis testing in various contexts such as sparse normal mixtures (Donoho and Jin, 2004; Cai et al., 2011), Gaussian sequence models (Ingster and Suslina, 2003), and correlated multivariate normal problems (Hall and Jin, 2010). However, very little work has been done on detection boundaries in generalized linear models for discrete outcomes.

In this paper, we study the detection boundary for hypothesis testing in the context of high-dimensional, sparse binary regression models. Motivated by case-control sequencing association studies for detecting the effects of rare variants on disease risk (Tang et al., 2014; Lee et al., 2014), we are interested in the complexity of the hypothesis testing problem when the design matrix is sparse. Specifically, rare variants are commonly observed in sequencing data. For example, in the Dallas Heart candidate gene sequencing study (Victor et al., 2004), 3476 individuals were sequenced in the region consisting of three genes ANGPTL3, ANGPTL4, and ANGPTL5. The goal of study was to test the effects of these genes on the risk of hypertriglyceridemia. A total of 93 genetic variants were observed in these genes. Each variant took values 0, 1, 2, which represents the number of minor alleles in a genetic variant. About half of the variants were singletons, *i.e.*, they were observed in only one person; 92 variants have the minor allele frequencies $< 5\%$. The design matrix is hence very sparse, with a vast majority of its columns having $< 5\%$ non-zero values (1 or 2), and the proportion of total non-zero elements in the design ma-

trix being $< 2.5\%$. It is expected only a small number of variants might be associated with hypertriglyceridemia. The presence of the sparse design matrix and sparse signals for binary outcomes results in substantial challenges in testing the association of these genes and hypertriglyceridemia.

Suppose there are n samples of binary outcomes, p covariates for each. Consider a binary regression model linking the outcomes to the covariates. We are interested in testing a global null hypothesis that the regression coefficients are all zero and the alternative is sparse with k signals, where $k = p^{1-\alpha}$ and $\alpha \in [0, 1]$. For binary regression models, we observe a new phenomenon in the behavior of detection boundaries which does not occur in the Gaussian framework, as explained below.

The main contribution of our paper is to derive the detection boundary for binary regression models as a function of two components: a notion of interaction of sparsity structure of the design matrix with the sparsity of the alternative and the minimal signal strength required for successful detection. Throughout we will call this notion of interaction between the sparsity structure of the design matrix and the sparsity of the alternative as the "sparsity interaction parameter" of the design matrix. This is unlike the results in Gaussian linear regression which has a one component detection boundary, namely the necessary signal strength. In the Gaussian linear model framework, Arias-Castro et al. (2011) and Ingster et al. (2010) show that if the design matrix satisfies certain 'low coherence conditions', then it is possible to detect the presence of a signal in a global sense, provided the signal exceeds a certain threshold in strength. In contrast, our results suggest that for binary regression problems, the difficulty of the problem is also determined by the sparsity interaction parameter of the design matrix. In this paper, we explore two key implications of this phenomenon which are outlined below.

First, if the sparsity interaction parameter of the design matrix is too high, we show that no signal can be detected irrespective of its strength. In Section 2.3, we provide sufficient conditions on the sparsity interaction parameter of the design matrix which yield such non-detectability problems. Such conditions on the sparsity interaction parameter corresponds to the first component of the detection boundary. Plan and Vershynin (2013a,b) discussed a difficulty in inference similar to that of ours, for design matrices with binary

entries in the context of 1-bit compressive sensing and sparse logistic models. Our results in Section 2.3 pertain to sparse design matrices with *arbitrary entries*, which are not necessarily orthogonal. We give a few examples of design matrices which satisfy our criteria for non-detectability. These include block diagonal matrices and banded matrices.

Second, for design matrices with binary entries and with low correlation among the columns, we are able to characterize both components of the detection boundary. In particular, if the sparsity interaction parameter of the design matrix is above a specified threshold, no signal is detectable irrespective of strength. Once again, this constitutes the first component of the detection boundary. Once the sparsity interaction parameter is below the same threshold, we also obtain the optimal thresholds with respect to the second component of the detection boundary, *i.e.*, the minimum signal strength required for successful detection. In this regime, our results parallel the theory of detection boundary in Gaussian linear regression. We also provide relevant tests to attain the optimal detection boundaries. In the sparse regime ($\alpha > \frac{1}{2}$), our results are sharp and rate adaptive in terms of the signal strength component of the detection boundary. Moreover, we observe a phase transition in both components of the detection boundary depending on the sparsity (α) of the alternative. To the best of our knowledge, this is the first work optimally characterizing a two component detection boundary in global testing problems against sparse alternatives.

To illustrate further, we contrast our results with the existing literature. In the case of a balanced one-way ANOVA type design matrix with each treatment having r independent replicates, for Gaussian linear models, Arias-Castro et al. (2011) show that the detection boundary is given by $O(\frac{p^{\frac{1}{4}}}{\sqrt{kr}})$ when $k \gtrsim \sqrt{p}$ and equals $\sqrt{\frac{2\rho^*_{\text{linear}}(\alpha) log(p)}{r}}$ when $k \ll \sqrt{p}$, where

$$\rho^*_{\text{linear}}(\alpha) = \begin{cases} \alpha - \frac{1}{2} & \text{if } \frac{1}{2} < \alpha < \frac{3}{4}, \\ (1 - \sqrt{1-\alpha})^2 & \text{if } \alpha \geq \frac{3}{4} \end{cases} \tag{2.1}$$

and $\rho^*_{\text{linear}}(\alpha)$ matches the detection boundary in Donoho and Jin (2004) in the normal mixture problem. For binary regression, we show that the detection boundary is drastically different and depends on the value of r. In particular, the sparsity sparsity interaction

parameter of the corresponding design matrix is given by $1/r$. For $r = 1$, every test is powerless irrespective of how strong the signal strength is under the alternative hypothesis. When $r > 1$, the behavior of the detection boundary can be categorized into three regimes. In the *dense* regime where $r > 1$ and $\alpha \leq \frac{1}{2}$, the detection boundary matches that of the Gaussian case up to rates and the usual generalized likelihood ratio test achieves the detection boundary. In the *sparse* regime, *i.e.*, when $\alpha > \frac{1}{2}$, the detection boundary behaves differently for $r \ll log(p)$ and $r \gg log(p)$. For $\alpha > \frac{1}{2}$ and $r \ll log(p)$, a new phenomenon arises: all tests are asymptotically powerless irrespective of how strong the signal strength is in the alternative. For $\alpha > \frac{1}{2}$ and $r \gg log(p)$, our results are identical to the Gaussian case, up to a constant factor accounting for the Fisher information. In this regime, we construct a version of the Higher Criticism test and show that this test achieves the lower bound. Despite the apparent simplicity of the balanced multiway design matrix studied in this paper, it presents significant challenges and exhibits interesting behavior. For the sparse case, we use the strong embedding theorem (Komlós et al., 1975) to obtain sharp detection boundary. Noting that this problem can also be cast as a test of homogeneity among p binomial populations with contamination in k of them, we also provide the corresponding detection boundary. Hence, roughly speaking, the two component detection boundary in our problem equals $(1, O(\frac{p^{\frac{1}{4}}}{\sqrt{kr}}))$ in dense regimes and $(O(\frac{1}{log(p)}), O(\sqrt{\frac{log(p)}{r}}))$ in sparse regimes, where the first component comprises of the order of $1/r$ or the sparsity interaction parameter and the second component indicates the order of signal strength required for successful detection.

Borrowing ideas from orthogonal designs, we obtain analogous results for binary design matrices which are sparse and have weak correlation among columns. Once again we are able to completely characterize the two component detection boundary in both dense and sparse regimes. Our versions of generalized likelihood ratio test and the Higher Criticism test continue to attain the optimal detection boundaries in dense and sparse regimes respectively. Similar to orthogonal designs, our results are sharp in the sparse regime and we once again obtain optimal phase transition in the two component detection boundary depending on the sparsity (α) of the alternative. In particular, our results show that under certain low correlation structures, the problem essentially behaves as an orthogonal

problem.

The rest of the paper is organized as follows. We first formally introduce our model in Section 2.2 and discuss general strategies. Here, we also provide a set of notations to be used throughout the paper. In Section 2.3, we study the non-detectability for sparse design matrices with arbitrary entries. In Section 2.4, we formally introduce a class of designs for which we will derive the sharp detection boundaries, namely, the Weak One Way ANOVA and the Strong One Way ANOVA designs. Section 2.5 introduces the Generalized Likelihood Ratio Test (GLRT) and the Higher Criticism Test in our designs, which will be used in subsequent sections to attain the sharp detection boundaries in two different regimes of sparsity. In Section 2.6, we first analyze the Strong One Way ANOVA designs and derive the sharp detection boundary in different sparsity regimes. In Section 2.7, we borrow intuition from Strong One Way ANOVA designs to derive the sharp detection boundary in different sparsity regimes for the Weak One Way ANOVA designs. Section 2.8 presents simulation studies which validate our theoretical results. Finally we collect all the technical proofs in Appendix B.

2.2 Preliminaries

Suppose there are n binary observations $y_i \in \{0,1\}$, for $1 \leq i \leq n$, with covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t$. The design matrix with rows $\boldsymbol{x}_i^t$ will be denoted by $\mathbf{X}$. Set $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$. The conditional distribution of y_i given $\boldsymbol{x}_i$ is given by

$$\mathbb{P}(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{\beta}) = \theta(\boldsymbol{x}_i^t \boldsymbol{\beta}) \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t \in \mathbb{R}^p$ is an unknown p-dimensional vector of regression coefficients. Henceforth, we will assume that θ is an arbitrary distribution function that is symmetric around 0, *i.e.*,

$$\theta(z) + \theta(-z) = 1 \text{ for all } z \in \mathbb{R}. \tag{2.3}$$

For some of the results, we will also require certain smoothness assumptions on $\theta(\cdot)$ which we will state when and where required. Examples of such $\theta(\cdot)$ include logistic and normal

distributions which respectively correspond to logistic and probit regression models. Let $M(\boldsymbol{\beta}) = \sum_{j=1}^{p} I(\beta_j \neq 0)$ and let $R_k^p = \{\boldsymbol{\beta} \in \mathbb{R}^p : M(\boldsymbol{\beta}) = k\}$. For some $A > 0$, we are interested in testing the global null hypothesis

$$H_0 : \boldsymbol{\beta} = 0 \text{ vs } H_1 : \boldsymbol{\beta} \in \Theta_k^A = \{\boldsymbol{\beta} \in \bigcup_{k' \geq k} R_{k'}^p : \min\{|\beta_j| : \beta_j \neq 0\} \geq A\}. \tag{2.4}$$

Set $k = p^{1-\alpha}$ with $\alpha \in (0, 1]$. We note that these types of alternatives has been considered by Arias-Castro et al. (2011), referred to as the "*Sparse Fixed Effects Model*" or SFEM. In particular, under the alternative, $\boldsymbol{\beta}$ has at least k non-zero coefficients exceeding A in absolute values. Alternatives corresponding to $\alpha \leq \frac{1}{2}$ belong to the *dense regime* and those corresponding to $\alpha > \frac{1}{2}$ belong to the *sparse regime*. We will denote by π a prior distribution on $\Theta_k^A \subset \mathbb{R}^p$. Throughout we will refer to A as the signal strength corresponding to the alternative in Equation (2.4).

We first recall a few familiar concepts from statistical decision theory. Let a test be a measurable function of the data taking values in $\{0, 1\}$. The Bayes risk of a test $\mathsf{T} = \mathsf{T}(\mathbf{X}, \mathbf{y})$ for testing $H_0 : \boldsymbol{\beta} = 0$ versus $H_1 : \boldsymbol{\beta} \sim \pi$ when H_0 and H_1 occur with the same probability, is defined as the sum of its probability of type I error (false positives) and its average probability of type II error (missed detection):

$$\text{Risk}_\pi(\mathsf{T}) := \mathbb{P}_0(\mathsf{T} = 1) + \pi[\mathbb{P}_{\boldsymbol{\beta}}(\mathsf{T} = 0)],$$

where $\mathbb{P}_{\boldsymbol{\beta}}$ denotes the probability distribution of $\mathbf{y}$ under model (2.2) and $\pi[\cdot]$ is the expectation with respect to the prior π. We study the asymptotic properties of the binary regression model (2.2) in the high-dimensional regime, *i.e.*, with $p \to \infty$ and $n = n(p) \to \infty$ and a sequence of priors $\{\pi_p\}$. Adopting the terminology from Arias-Castro et al. (2011), we say that a sequence of tests $\{\mathsf{T}_{n,p}\}$ is *asymptotically powerful* if $\lim_{p\to\infty} \text{Risk}_{\pi_p}(\mathsf{T}_{n,p}) = 0$, and it is *asymptotically powerless* if $\liminf_{p\to\infty} \text{Risk}_{\pi_p}(\mathsf{T}_{n,p}) \geq 1$. When no prior is specified, the risk is understood to be the worst case risk or the minimax risk defined as

$$\text{Risk}(\mathsf{T}) := \mathbb{P}_0(\mathsf{T} = 1) + \max_{\boldsymbol{\beta} \in \Theta_k^A} [\mathbb{P}_{\boldsymbol{\beta}}(\mathsf{T} = 0)] \, .$$

The detection boundary of the testing problem (2.4) is the demarcation of signal strength A which determines whether all tests are asymptotically powerless (we call this Lower Bound of the problem) or there exists some test which is asymptotically powerful (we call this the Upper Bound of the problem).

To understand the minimax risk, set

$$d(\mathcal{P}_0, \mathcal{P}_1) = \inf\{|P - Q|_1 : P \in \mathcal{P}_0, Q \in \mathcal{P}_1\},$$

where $\mathcal{P}_0, \mathcal{P}_1$ are two families of probability measures and $|P - Q|_1 = \sup_B |P(A) - Q(A)|$, with B being a Borel set in $\mathbb{R}^n$, denotes the total-variation norm. Then for any test T, we have (Wald, 1950)

$$\mathrm{Risk}(\mathsf{T}) \geq 1 - \tfrac{1}{2} d(\mathbb{P}_0, \mathrm{conv}_{\boldsymbol{\beta} \in \Theta_k^A}(\mathbb{P}_{\boldsymbol{\beta}})),$$

where conv denotes the convex hull. However, $d(\mathbb{P}_0, \mathrm{conv}_{\boldsymbol{\beta} \in \Theta_k^A}(\mathbb{P}_{\boldsymbol{\beta}}))$ is difficult to calculate. But it is easy to see that for any test T and any prior π, one has $\mathrm{Risk}(\mathsf{T}) \geq \mathrm{Risk}_\pi(\mathsf{T})$. So in order to prove that a sequence of tests is asymptotically powerful, it suffices to bound from above the worst-case risk $\mathrm{Risk}(\mathsf{T})$. Similarly, in order to show that all tests are asymptotically powerless, it suffices to work with an appropriate prior to make calculations easier and bound the corresponding risk from below for any test T.

It is worth noting that, for any prior π on the set of k-sparse vectors in $\mathbb{R}^p$ and for any test T, we have

$$\mathrm{Risk}_\pi(\mathsf{T}) \geq 1 - \tfrac{1}{2}\mathbb{E}_0|L_\pi - 1| \geq 1 - \tfrac{1}{2}\sqrt{\mathbb{E}_0(L_\pi^2) - 1},$$

where L_π is the π-integrated likelihood ratio and $\mathbb{E}_0$ denotes the expectation under H_0. For the model (2.2), we have

$$L_\pi = 2^n \int \prod_{i=1}^n \left(\frac{\theta(\boldsymbol{x}_i^t \boldsymbol{\beta})}{\theta(-\boldsymbol{x}_i^t \boldsymbol{\beta})} \right)^{y_i} \theta(-\boldsymbol{x}_i^t \boldsymbol{\beta}) d\pi(\boldsymbol{\beta}). \tag{2.5}$$

Hence in order to assess the lower bound for the risk, it suffices to bound from above $\mathbb{E}_0(L_\pi^2)$. By Fubini's theorem, for fixed design matrix $\mathbf{X}$, we have

$$\mathbb{E}_0(L_\pi^2) = 2^n \iint \prod_{j=1}^n \left[\theta(\boldsymbol{x}_i^t \boldsymbol{\beta}) \theta(\boldsymbol{x}_i^t \boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t \boldsymbol{\beta}) \theta(-\boldsymbol{x}_i^t \boldsymbol{\beta}') \right] d\pi(\boldsymbol{\beta}) d\pi(\boldsymbol{\beta}'), \tag{2.6}$$

where $\boldsymbol{\beta}, \boldsymbol{\beta}^{'} \sim \pi$ are independent. In the rest of the paper, all of our analysis is based on studying $\mathbb{E}_0(L_\pi^2)$ carefully for the prior distribution π chosen below.

In the context of finding an appropriate test matching the lower bound, by the Neyman-Pearson Lemma, the test which rejects when $L_\pi > 1$ is the most powerful Bayes test and has risk equal to $1 - \frac{1}{2}\mathbb{E}_0|L_\pi - 1|$. However this test requires knowledge of the sparsity index α and is also computationally intensive. Hence we will construct tests which do not require knowledge of α and are computationally much less cumbersome.

Ideally, one seeks least favorable priors, *i.e.*, those priors for which the minimum Bayes risk equals the minimax risk. Inspired by Baraud (2002), we choose π to be uniform over all k sparse subsets of $\mathbb{R}^p$ with signal strength either A or $-A$.

2.2.1 Notations

We provide a brief summary of notation used in the paper. For two sequences of real numbers a_p and b_p, we say $a_p \ll b_p$ or $a_p = o(b_p)$, when $\limsup_{p\to\infty} \frac{a_p}{b_p} \to 0$ and we say $a_p \lesssim b_p$ or $a_p = O(b_p)$ if $\limsup_{p\to\infty} \frac{a_p}{b_p} < \infty$. The indicator function of a set B will be denoted by $\mathbf{I}(B)$. We take π to be uniform over all k sparse subsets of $\mathbb{R}^p$ with signal strength either A or $-A$. Let $M(k,p)$ be the collection of all subsets of $\{1,\ldots,p\}$ of size k. For each $m \in M(k,p)$, let $\xi^m = (\xi_j)_{j\in m}$ be a sequence of independent Rademacher random variables taking values in $\{+1,-1\}$ with equal probability. Given $A > 0$ for testing (2.4), a realization from the prior distribution π on $\mathbb{R}^p$ can be expressed as

$$\beta_{\xi,m} = \sum_{j\in m} A\xi_j e_j,$$

where $(e_j)_{j=1}^p$ is the canonical basis of $\mathbb{R}^p$ and m is uniformly chosen from $M(k,p)$. Since, the alternative in (2.4) allows both positive and negative directions of signal strength β_j, we call it a two-sided alternative. On the contrary, when we are given the extra information in (2.4) that the β_j's have the same sign, then we call the alternative a one-sided alternative. A realization from a prior distribution over one-sided k sparse alternatives can be expressed as $\sum_{j\in m} A\xi e_j$, where ξ is a single Rademacher random variable.

For any distribution $\pi^{'}$ on $M(k,p)$, by **support**$(\pi^{'})$ we mean the smallest set $I^{'} := \{M :$

$M \in M(k,p)\}$ such that $\pi^{'}(I^{'}) = 1$. For any distribution π^* over $M(k,p)$, we say that another distribution π_0 over $M(k,p)$ is equivalent to π^* (denoted by $\pi_0 \sim \pi^*$) if π_0 is uniform on its support and

$$\pi^*(M \notin \textbf{support}(\pi_0)) = o(1).$$

By the support of a vector $v \in \mathbb{R}^p$, we mean the set $\{j \in \{1,\ldots,p\} : v_j \neq 0\}$; the vector v is Q-sparse if the support of v has at most Q elements. For $i = 1\ldots,n$, we will denote the support of the i^{th} row of $\boldsymbol{X}$ by $S_i := \{j : \boldsymbol{X}_{i,j} \neq 0\} \subset \{1,\ldots,p\}$. Let BC^l denote the set of all functions whose l^{th} derivative is continuous and bounded over $\mathbb{R}$. By $\theta(\cdot) \in \text{BC}^l(0)$, we mean that the l^{th} derivative of $\theta(\cdot)$ is continuous and bounded in a neighborhood of 0. Finally, by saying that a sequence measurable map $\chi_{n,p}(y, \boldsymbol{X})$ of the data is tight, we mean that it is stochastically bounded as $n, p \to \infty$.

2.3 Sparse Design Matrices and Non-detectability of Signals

In this section, we study the effects of sparsity structures of the design matrix $\boldsymbol{X}$ on the detection of signals. Our key results in Theorem 2.1 below provide a sufficient condition on the sparsity structure of the $\boldsymbol{X}$ which renders all tests asymptotically powerless in the sparse regime irrespective of signal strength A. This result for non-detectability is quite general and are satisfied by different classes of sparse design matrices as we discuss below. We verify the hypothesis of Theorem 2.1 in a few instances where certain global detection problems can be extremely difficult.

Let $\pi_0 \sim \pi$ and R_{π_0} denote the support of π_0. For a sequence of positive integers σ_p, we say that $j_1, j_2 \in \{1,\ldots,p\}$ are "σ_p-mutually close" if $|j_1 - j_2| \leq \sigma_p$. For an m_1 from π_0 and $N \geq 0$, let $R^N_{m_1}(\sigma_p)$ denote the set of all $\{l_1,\ldots,l_k\} \in R_{\pi_0}$ such that there are exactly N elements "σ_p-mutually close" with members of m_1.

Theorem 2.1. *Let $k = p^{1-\alpha}$ with $\alpha > \frac{1}{2}$. Let $\pi_0 \sim \pi$ and $\{\sigma_p\}$ be a sequence of positive integers with $\sigma_p \ll p^{\epsilon}$ for all $\epsilon > 0$. Let m_1 be drawn from π_0. Suppose that for all $N = 0,\ldots,k$ and every*

m_2 drawn from π_0 with $m_2 \in R^N_{m_1}(\sigma_p)$, the following holds for some sequence $\delta_p > 0$:

$$\sum_{i=1}^{n} \{\mathbf{I}(\min\{|m_1 \cap S_i|, |m_2 \cap S_i|\} > 0)\} \leq N\delta_p, \tag{2.7}$$

where S_i is defined in the last paragraph of Section 2. Then if $\delta_p \ll log(p)$, all tests are asymptotically powerless.

An intuitive explanation of Theorem 2.1 is as follows. If the support of $\boldsymbol{\beta}$ under the alternative does not intersect the support of a row of the design matrix $\boldsymbol{X}$, the observation corresponding to that particular row does not provide any information about the alternative hypothesis. If randomly selected draws from $M(k, p)$ fail to intersect with the support of most of the rows, as quantified by Equation (2.7), then all tests will be asymptotically powerless irrespective of the signal strength in the alternative. Also intuitively, the quantity $\frac{1}{\delta_p}$ in Theorem 2.1 is the candidate for sparsity interaction parameter of $\boldsymbol{X}$ since if $\frac{1}{\delta_p}$ is too large, as quantified by $\frac{1}{\delta_p} \gg \frac{1}{log(p)}$, then all tests are asymptotically powerless in the sparse regime irrespective of the signal strength. Now we provide a few examples where condition (2.7) can be verified to hold for appropriate parameters.

Example 1: Block Structure

Suppose that, up to permutation of rows, $\boldsymbol{X}$ can be partitioned into a block diagonal matrix consisting of $\mathbf{G}^{(1)}, \ldots, \mathbf{G}^{(M)}$ and a matrix $\mathbf{G}$ as follows:

$$\mathbf{X} = \left[\begin{array}{cccc} \mathbf{G}^{(1)}_{c_1 \times d_1} & & & \\ & \ddots & & \\ & & \mathbf{G}^{(j)}_{c_j \times d_j} & \\ & & & \ddots \\ & & & \mathbf{G}^{(M)}_{c_M \times d_M} \\ \hline & & \mathbf{G}_{\tilde{c} \times p} & \end{array}\right] \in \mathbb{R}^{n \times p} \tag{2.8}$$

where $\tilde{c} = n - \sum_{j=1}^{M} c_j$. The matrices $\mathbf{G}, \mathbf{G}^{(1)}, \ldots, \mathbf{G}^{(M)}$ are *arbitrary matrices* of specified dimensions. Let $c^* = \max_{1 \leq j \leq M} c_j$ and $l^* = \max_{1 \leq j \leq M} d_j$. Indeed c^*, l^* and the structure of $\mathbf{G}$

decide the sparsity of the design matrix $\boldsymbol{X}$. In Theorem 2.2 below, we provide necessary conditions on c^*, l^* and $\mathbf{G}$ which dictate the validity of condition (2.7) and hence renders all tests asymptotically powerless irrespective of signal strength.

Theorem 2.2. *Assume that the matrix $\boldsymbol{X}$ is of the form given by* (2.8)*. Let $k = p^{1-\alpha}$ with $\alpha > \frac{1}{2}$ and suppose that $|\bigcup_{i>n^*} S_i| \ll p$ where $n^* = \sum_{j=1}^{M} c_j$. Let $l^* \ll p^{\epsilon}$ for all $\epsilon > 0$. If $c^* \ll log(p)$, then condition* (2.7) *holds and thus all tests are asymptotically powerless.*

In Theorem 2.2, the condition $|\bigcup_{i>n^*} S_i| \leq\ll p$ is an assumption on the structure of $\mathbf{G}$ which restricts the locations of non-zero elements of $\mathbf{G}$. This condition on $\mathbf{G}$ is not tight and can be much relaxed provided one assumes further structures on $\mathbf{G}$. In effect, this implies that asymptotically the bulk of the information about the alternatives comes from the block diagonal part of $\boldsymbol{X}$ and the information from $\mathbf{G}$ is asymptotically negligible.
Also intuitively, $\frac{1}{c^*}$ is the candidate for the sparsity interaction parameter since if $\frac{1}{c^*}$ is too high, as quantified by $\frac{1}{c^*} \gg \frac{1}{log(p)}$, then all tests are asymptotically powerless in the sparse regime. It is natural to ask about the situation when the sparsity interaction parameter is below the specified threshold of $\frac{1}{log(p)}$, *i.e.*, $c^* \gg log(p)$. To this end, it is possible to analyze the necessary and sufficient conditions on the signal strength A dictating asymptotic detectability in problem (2.4) when $c^* \gg log(p)$ for $\boldsymbol{X}$ in (2.8) but possibly with $|\bigcup_{i>n^*} S_i| \gg p$. In Section 2.7, we provide an answer to this question when $\boldsymbol{X}$ has binary entries.

Example 2: Banded Matrix

Suppose $\boldsymbol{X}$ has the following banded structure, possibly after a permutation of its rows. Suppose there exists $l_2 > l_1$ such that for $i = 1, \ldots, n$, $X_{i,j} = 0$ for $j < i - l_1$ or $j > i + l_2$. Further, let $|\bigcup_{i>n} S_i| \ll p$. Note that this allows design matrices $\boldsymbol{X}$ which can be partitioned into a banded matrix of band-width $l_2 - l_1$ and an arbitrary design matrix with sparsity restrictions as specified by $|\bigcup_{i>n} S_i| \ll p$.

Theorem 2.3. *Let $k = p^{1-\alpha}$ with $\alpha > \frac{1}{2}$. Suppose $\boldsymbol{X}$ is a banded design matrix as described above. Suppose that $l_2 - l_1 \ll log(p)$. Then condition* (2.7) *holds and thus all tests are asymptotically powerless.*

2.4 ANOVA-type Design Matrices

In Section 2.3, we provided conditions on $\boldsymbol{X}$ under which all tests are asymptotically powerless irrespective of signal strength A. To complement those results, the subsequent sections will be devoted towards analyzing situations when $\boldsymbol{X}$ is not pathologically sparse and hence one can expect to study non-trivial conditions on the signal strength A that determine the complexity in (2.4). In this section we introduce certain design matrices with binary entries motivated by sequencing association studies. In subsequent sections we will derive the detection boundary for binary regression models with these design matrices.

In order to introduce the design matrices we wish to study, we need some notations. Set $\Omega^* = \{i : |S_i| = 1\}$. For $j = 1, \ldots, p$, let $\Omega_j^* = \{i \in \Omega^* : S_i = \{j\}\}$ with $r_j = |\{i \in \Omega^* : S_i = \{j\}\}|$. Let $r^* = \max_{1 \le j \le p} r_j$ and $r_* = \min_{1 \le j \le p} r_j$. Also, let $n^* = \sum_{j=1}^p r_j$ and $n_* = n - n^*$. In words, for each j, Ω_j^* is the collection of individuals with only one non-zero informative covariate appearing as the j^{th} covariate and r_j is the number of such individuals.

Definition 2.1. *We say that the design matrix $\boldsymbol{X}$ is a Weak One Way ANOVA (*WA*) design and denote it by $\boldsymbol{X} \in \mathrm{WA} = \mathrm{WA}(n^*, n_*, r^*, r_*, Q_{n,p}, \gamma_{n,p})$ if the following conditions hold:*

C1: *The design matrix $\boldsymbol{X}_{n \times p}$ has binary entries;*

C2: *$|S_i| \le Q_{n,p}$ for all $i = 1 \ldots, n$, for some sequence $Q_{n,p}$;*

C3: *$\frac{n_* Q_{n,p}^2}{r^*} \ll \gamma_{n,p}$ for some sequence $\gamma_{n,p} \to \infty$.*

As a special case of the above definition, we have the following definition.

Definition 2.2. *We say that the design matrix $\boldsymbol{X}$ is a Strong One Way ANOVA (*SA*) design, and denote it by $\boldsymbol{X} \in \mathrm{SA}(r)$, if it is a* WA *design with $r_* = r^* = r$ and $n_* = 0$.*

A few comments are in order for the above set of assumptions in Definitions 2.1 and 2.2. The motivation for condition **C1** comes from genetic association studies assuming a dominant model. As our proofs will suggest, this can be easily relaxed, allowing the elements of $\boldsymbol{X}$ to be uniformly bounded above and below. Condition **C2** imposes sparsity on $\boldsymbol{X}$. Finally, since the part of $\boldsymbol{X}$ without $\mathbf{G}$ is exactly orthogonal, condition **C3** restricts the deviation of $\boldsymbol{X}$ from exact orthogonality. In particular, if the size of $\mathbf{G}$ is "not too

large" compared to orthogonal part of $\boldsymbol{X}$, as we will quantify later, then the behavior of the detection problem is similar to the one with an exactly orthogonal design. In essence, this captures low correlation designs suitable for binary regression with ideas similar to low coherence designs as imposed by Arias-Castro et al. (2011) for Gaussian linear regression.

A binary design matrix is orthogonal if and only if all of its rows have at most one non-zero element. Hence, up to a permutation of rows, any binary design matrix can be potentially partitioned as a one-way balanced ANOVA design and an arbitrary matrix. In particular, up to a permutation of rows, any binary design matrix is equivalent to Equation (2.8) where each $\mathbf{G}^{(j)}_{r_j \times 1} = (1, \ldots, 1)^t$, $c_j = r_j$, $d_j = 1$, $c^* = r^*$, $l^* = 1$, $\tilde{c} = n_*$ and $\mathbf{G}$ is an arbitrary matrix with binary entries. Because of the presence of $\mathbf{G}$, WA designs allow for correlated binary design matrices with sparse structures. However, condition **C3** restricts the size of $\mathbf{G}$ (numerator) compared to the orthogonal part (denominator) by a factor of $\gamma_{n,p}$. Intuitively, this implies low correlation structures in $\boldsymbol{X}$. The condition **C3** restricts the effect of $\mathbf{G}$ on the correlation structures of $\boldsymbol{X}$ by not allowing too many rows compared to the size of the orthogonal part of $\boldsymbol{X}$. It is easy to see that when $n_* Q_p \ll p$, then since $|\bigcup_{i \notin \Omega^*} S_i| \ll p$, one can essentially ignore the rows outside Ω^* using an argument similar to that in the proof of Theorem 2.2 and the problem becomes equivalent to $\mathrm{SA}(r_*)$ designs. However, condition **C3** allows for the cases $|\bigcup_{i \notin \Omega^*} S_i| \gg p$. For example, if $Q = log(p)^b$ for some $b > 0$, then as long as $r^* \gamma_p \gg p a_p log(p)^b$ for some sequence $a_p \to \infty$, one can potentially have $n_* Q_p \gg p$ and hence the simple reduction of the problem as in proof of Theorem 2.2 is no longer possible. In order to show that the detection problem still behaves similar to an orthogonal design, one needs much subtler analysis to ignore the information about the alternative coming from the subjects corresponding to $\mathbf{G}$ part of the design $\boldsymbol{X}$. Therefore, condition **C3** allows for a rich class of correlation structures in $\boldsymbol{X}$.

As mentioned earlier, a major motivation of study comes from an effort to understand rare variants sequence data. In particular, recall the Dallas Heart Study described in the introduction. In Figure 2.1 we present a heat map of the genotype matrix of the first 500

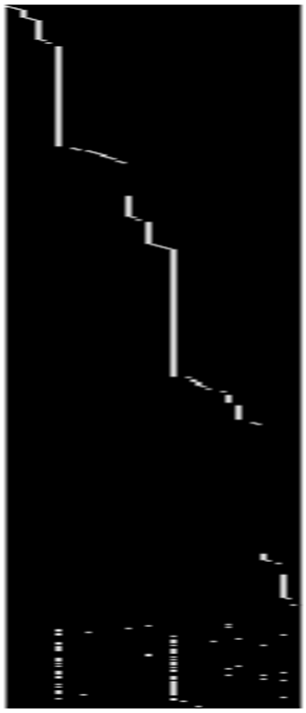

Figure 2.1: Heat map of genotype matrix of first 500 subjects of Dallas Heart Study data after suitable rearrangement of subject indices.

subjects of the data after suitable rearrangement of subject indices with the 53rd or the common variant removed. The non-zero entries of the genotype matrix corresponding to location of mutation have been colored black while the zero entries corresponding to no mutation are colored in white. Interestingly, the genotype matrix can be seen to be partitioned into two parts. The top of the matrix is an orthogonal block diagonal structure similar to $\boldsymbol{X}$ described above and the bottom part is a non-orthogonal sparse matrix which corresponds to $\mathbf{G}$ in our definition. In subsequent sections, we study the role of the parameter vector
$(n^*, n_*, r^*, r_*, Q_{n,p}, \gamma_{n,p})$ in deciding the detection boundary. We first present the analysis of relatively simpler SA designs followed by the study of WA designs. The analysis of simpler SA designs provides the crux of insight for the study of detection boundary under low correlation WA designs, and at the same time yields cleaner results for easier interpretation. We will demonstrate that the quantity $\frac{1}{r}$ is the sparsity interaction parameter when $\boldsymbol{X} \in \mathrm{SA}(r)$. In the case of WA, r^* and r_* play the same role as that of r in $\mathrm{SA}(r)$

designs. We divide our study of each design into two main sections, namely the Dense Regime ($\alpha \leq \frac{1}{2}$) and the Sparse Regime ($\alpha > \frac{1}{2}$). In the next section, we first introduce the tests which will be essential for attaining the optimal detection boundaries in dense and sparse regimes respectively.

2.5 Tests

We propose in this section the generalized likelihood-ratio test and a higher-criticism test for binary regression models. We begin by defining Z-statistics for WA and $\mathrm{SA}(r)$ designs which will be required for introducing and analyzing upper bounds later. Also, in order to separate the information about the alternative coming from the $\mathbf{G}$ part of $\boldsymbol{X}$, we define a Z-statistic separately for the non-orthogonal part. With this in mind we have the following definitions.

Definition 2.3. *Let* $\boldsymbol{X} \in \mathrm{WA}$.

1. *Define the* j^{th} *Z-statistic as follows*

$$Z_j = \sum_{i \in \Omega_j} y_i, \ j = 1, \ldots, p.$$

2. *Letting* $\mathbf{G} = \{\mathbf{G}_{ij}\}_{n_* \times p}$ *define*

$$Z_j^{\mathbf{G}} = \sum_{i=n-n_*+1}^{n} \mathbf{G}_{ij} y_i, \ j = 1, \ldots, p$$

With these definitions we are now ready to construct our tests.

2.5.1 The Generalized Likelihood Ratio Test (GLRT)

We now introduce a test that will be used to attain the detection boundary in the dense regime. Let Z_j be the j^{th} Z-statistic in Definition 2.3. Then the Generalized Likelihood Ratio Test is based on the following test statistic:

$$\mathsf{T}_{\mathrm{GLRT}} := \sum_{j=1}^{p} \frac{4(Z_j - \frac{r_j}{2})^2}{r_j} \ . \tag{2.9}$$

Under H_0 we have, $\mathbb{E}_{H_0}(\mathsf{T}_{\mathrm{GLRT}}) = p$ and $\mathrm{Var}_{H_0}(\mathsf{T}_{\mathrm{GLRT}}) = O(p)$. Hence $\frac{\mathsf{T}_{\mathrm{GLRT}}-p}{\sqrt{2p}}$ is tight. Our test rejects when

$$\frac{\mathsf{T}_{\mathrm{GLRT}} - p}{\sqrt{2p}} > t_p$$

for a suitable t_p to be decided later.

Note that this test only uses partial information from the data. Since we shall show that, asymptotically using this partial information is sufficient, we will not lose power in an asymptotic sense. However, from finite sample performance point of view, it is more reasonable to use the following test by combining information from $\mathbf{G}$ which can be explained as a combination of GLRT based on the orthogonal and non-orthogonal parts of $\boldsymbol{X}$ respectively.

$$\text{Reject when : } \max\{\frac{\mathsf{T}_{\mathrm{GLRT}} - p}{\sqrt{2p}}, \frac{\sum_{j=1}^{p}\left[(Z_j^{\mathbf{G}})^2 - \mathbb{E}_{H_0}((Z_j^{\mathbf{G}})^2)\right]}{\sqrt{\mathrm{V}_{H_0}(\sum_{j=1}^{p}(Z_j^{\mathbf{G}})^2)}}\} > t_p$$

We note that given a particular $\mathbf{G}$, the quantities $\mathbb{E}_{H_0}((Z_j^{\mathbf{G}})^2)$ and $\mathrm{V}_{H_0}(\sum_{j=1}^{p}(Z_j^{\mathbf{G}})^2)$ can be easily calculated by simple moment calculations of Bernoulli random variables. We do not go into specific details here. Finally, since combining correct size tests by Bonferroni correction does not change asymptotic power, our proofs about asymptotic power continue to hold for this modified GLRT without any change.

2.5.2 Version of Higher Criticism Test

Assume $r_* \geq 2$. Let R_j be a generic $\mathrm{Bin}(r_j, \frac{1}{2})$ random variable and $\mathbb{B}_j, \overline{\mathbb{B}_j}$ respectively denote the distribution function and the survival function of $\frac{|R_j - \frac{r_j}{2}|}{\sqrt{\frac{r_j}{4}}}$. Hence

$$\mathbb{B}_j(t) = \mathbb{P}\Big(\frac{|R_j - \frac{r_j}{2}|}{\sqrt{\frac{r_j}{4}}} \leq t\Big), \overline{\mathbb{B}_j}(t) = 1 - \mathbb{B}_j(t)\ .$$

From Definition 2.3, the Z_j's are independent $\mathrm{Bin}(r_j, \frac{1}{2})$ under H_0 for $j = 1, \ldots, p$. Let

$$W_p(t) = \frac{\sum_{j=1}^{p}\mathbf{I}\Big(\frac{|Z_j - \frac{r_j}{2}|}{\sqrt{\frac{r_j}{4}}} > t\Big) - \overline{\mathbb{B}_j}(t)}{\sqrt{\sum_{j=1}^{p}\overline{\mathbb{B}_j}(t)(1 - \overline{\mathbb{B}_j}(t))}}.$$

Now we define the Higher Criticism Test as

$$\mathsf{T}_{\mathrm{HC}} := \max_{t \in [1, \sqrt{3log(p)}] \cap \mathbb{N}} W_p(t) \tag{2.10}$$

where $\mathbb{N}$ denotes the set of natural numbers. The next theorem provides the rejection region for the Higher Criticism Test.

Theorem 2.4. *For* WA *designs,* $\lim_{p\to\infty} \mathbb{P}_{H_0}(\mathsf{T}_{\mathrm{HC}} > log(p)) = 0$.

Hence one can use $(1+\epsilon)log(p)$ as a cutoff to construct a test based on T_{HC} for any arbitrary fixed $\epsilon > 0$:

$$\text{Higher Criticism Test} : \text{Reject when } \mathsf{T}_{\mathrm{HC}} > (1+\epsilon)log(p). \tag{2.11}$$

By Theorem 2.4, the above test based on T_{HC} has asymptotic type I error converging to 0. We note that, when $log(p) \ll r_*$, we can obtain a rejection region of the form $\mathsf{T}_{\mathrm{HC}} > \sqrt{2(1+\epsilon)loglog(p)}$ while maintaining asymptotic type I error control. This type of rejection region is common in the Higher Criticism literature. As we will see in Section 2.6, the interesting regime where the Higher Criticism Test is important is when $log(p) \ll r_*$. In this regime we can have the same rejection region of the Higher Criticism as obtained in Donoho and Jin (2004); Hall and Jin (2010). However, for generality we will instead work with the rejection region given by Equation (2.11).

Once again, note that this test only uses partial information from the data. Since we shall show that, asymptotically, using this partial information is sufficient, we will not lose power in an asymptotic sense. However, from a finite sample performance point of view it is more reasonable to use the following test by combining information from $\mathbf{G}$ which can be explained as a combination of Higher Criticism Tests based on the orthogonal and non-orthogonal parts of $\boldsymbol{X}$ respectively. Letting $g_j = \sum_{i>n^*} \boldsymbol{X}_{ij},\ j = 1, \ldots, p$, define the Higher Criticism type test statistic based on $\mathbf{G}$ as

$$W_p^{\mathbf{G}}(t) = \frac{\sum_{j=1}^{p} \mathbf{I}\Big(\frac{|Z_j^{\mathbf{G}} - \frac{g_j}{2}|}{\sqrt{\frac{g_j}{4}}} > t\Big) - \mathbb{P}_{H_0}\Big(\frac{|Z_j^{\mathbf{G}} - \frac{g_j}{2}|}{\sqrt{\frac{g_j}{4}}} > t\Big)}{\sqrt{\mathrm{Var}_{H_0} \sum_{j=1}^{p} \mathbf{I}\Big(\frac{|Z_j^{\mathbf{G}} - \frac{g_j}{2}|}{\sqrt{\frac{g_j}{4}}} > t\Big)}}.$$

The quantities $\mathbb{P}_{H_0}\Big(\frac{|Z_j^{\mathbf{G}}-\frac{g_j}{2}|}{\sqrt{\frac{g_j}{4}}} > t\Big)$ and $\mathrm{Var}_{H_0}\sum_{j=1}^{p}\mathbf{I}\Big(\frac{|Z_j^{\mathbf{G}}-\frac{g_j}{2}|}{\sqrt{\frac{g_j}{4}}} > t\Big)$ can be suitably approximated given particular instances of $\mathbf{G}$. However, we omit the specific details here for coherence of exposition. Finally, defining $W_p(t) = \max\{W_p(t), W_p^{\mathbf{G}}(t)\}$, one can follow the previous steps in defining the Higher Criticism Test with exactly similar arguments. Since combining correct size tests by Bonferroni correction does not change asymptotic power, the proofs concerning the power of the resulting test goes through with similar arguments. We omit the details here.

2.6 Detection Boundary and Asymptotic Analysis for SA Designs

We begin by noting that the $\mathrm{SA}(r)$ designs can be equivalently cast as a problem of testing homogeneity among p different binomial populations with r trials each. Suppose

$$y_j \sim \mathrm{Bin}(\mathrm{r}, \frac{1}{2} + \nu_{\mathrm{j}}) \text{ independent for } j = 1, \ldots, p. \tag{2.12}$$

Let $\nu = (\nu_1, \ldots, \nu_p)^t$. For some $\Delta \in (0, \frac{1}{2}]$, we are interested in testing the global null hypothesis

$$H_0 : \nu = 0 \text{ vs } H_1 : \nu \in \Xi_k^{\Delta} = \{\nu \in R_k^p : \min\{|\nu_j| : \nu_j \neq 0\} \geq \Delta\}. \tag{2.13}$$

When $\boldsymbol{X} \in \mathrm{SA}(r)$, the models (2.2) and (2.12) are equivalent with $\eta_j = \theta(\beta_j) - \frac{1}{2}$. Hence, sparsity in $\boldsymbol{\beta}$ is equivalent to sparsity in ν in the sense that $\boldsymbol{\beta} \in R_k^p$ if and only if $\nu \in R_k^p$. Further, rates of Δ which decide asymptotic detectability of (2.13) can be related to rates of A which determine detectability in (2.4) when the link function θ is continuously differentiable in a neighborhood around 0.

Remark 2.1. When θ is the distribution function for $\mathrm{U}(-\frac{1}{2}, \frac{1}{2})$, $\nu_j = \beta_j$ for all $j = 1, \ldots, p$. Hence, the detection boundary in problem (2.13) follows from that in problem (2.4) by taking θ to be the distribution function of $\mathrm{U}(-\frac{1}{2}, \frac{1}{2})$, *i.e.*, $\theta(x) = (x + \frac{1}{2})\,\mathbf{I}(-\frac{1}{2} < x < \frac{1}{2})$.

Remark 2.2. The prior π_{eq} that we will use for the binomial homogeneity of proportion testing is as follows. For each $m \in M(k, p)$, let $\xi^m = (\xi_j)_{j \in m}$ be a sequence of independent Rademacher random variables taking values in $\{+1, -1\}$ with equal probability. Given

$\Delta \in (0, \frac{1}{2})$ for testing (2.13), a realization from the prior distribution π_{eq} on $\mathbb{R}^p$ can be expressed as $\nu_{\xi,m} = \sum_{j \in m} \Delta \xi_j e_j$, where $(e_j)_{j=1}^p$ is the canonical basis of $\mathbb{R}^p$ and m is uniformly chosen from $M(k,p)$. Note that given the prior π on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^r$ discussed earlier, π_{eq} is the prior induced on $\nu = (\nu_1, \ldots, \nu_p)^t$ with $\frac{1}{2} + \nu_j = \theta(\beta_j)$ for $j = 1 \ldots, p$.

Owing to Remark 2.1, one can deduce the detection boundary of the binomial proportion model (2.12) from the detection boundary in $\mathrm{SA}(r)$ designs. However, for the sake of easy reference, we provide the detection boundaries for both models. Before proceeding further, we first state a simple result about SA designs, a part of which directly follows from Theorem 2.1. Note that $\mathrm{SA}(1)$ design corresponds to the case when the design matrix is identity $I_{p\times p}$. Unlike Gaussian linear models, for binary regression, when the design matrix is identity, for two-sided alternatives, all tests are asymptotically powerless irrespective of sparsity (*i.e.*, in both dense and sparse regimes) and signal strengths. Such a result arises for $r = 1$ because we allow the alternative to be two-sided.

In the modified problem where one only considers the one-sided alternatives, all tests still remain asymptotically powerless irrespective of signal strengths in the sparse regime, *i.e.*, when $\alpha > \frac{1}{2}$. However in the dense regime, *i.e.*, when $\alpha \leq \frac{1}{2}$, the problem becomes non-trivial and the test based on the total number of successes attains the detection boundary. The detection boundary for this particular problem is provided in Theorem 2.5 part 2(b). Also, in the one-sided problem, the Bayes Test can be explicitly evaluated and quite intuitively turns out to be a function of the total number of successes. In the next theorem, we collect all these results. The proof of the Theorem can be found in the supplementary material.

Theorem 2.5. *Assume $\boldsymbol{X} \in \mathrm{SA}(1)$, which assumes $r = 1$ and $\boldsymbol{X} = I$. Then the following holds for both the problems* (2.4) *and* (2.13)*.*

1. *For two-sided alternatives all tests are asymptotically powerless irrespective of sparsity and signal strength.*

2. *For one-sided alternatives :*

   (a) suppose $\theta \in \mathrm{BC}^1(0)$. Then in the dense regime ($\alpha \leq \frac{1}{2}$), all tests are asymptotically

powerless if $\frac{A^2}{p^{1-2\alpha}} \to 0$ *in problem* (2.4) *or* $\frac{\Delta^2}{p^{1-2\alpha}} \to 0$ *in problem* (2.13). *Further if* $\frac{A^2}{p^{1-2\alpha}} \to \infty$ *in problem* (2.4) *or* $\frac{\Delta^2}{p^{1-2\alpha}} \to \infty$ *in problem* (2.13), *then the test based on the total number of successes* ($\sum_{i=1}^{p} y_i$) *is asymptotically powerful.*

(b) in sparse regime ($\alpha > \frac{1}{2}$), *all tests are asymptotically powerless.*

The case of two-sided of alternatives when $r = 1$ can indeed be understood in the following way. Under the null hypothesis, each y_i is an independent Bernoulli$(1/2)$ random variable and under the prior on the alternative which allows each β_i to be $+A$ or $-A$ with probability $\frac{1}{2}$, the y_i's are again independent Bernoulli$(1/2)$ random variables. So of course there is no way to distinguish them based on the observations y_i's when the $\boldsymbol{\beta}$ is generated according to the prior mentioned earlier. Our proof is based on this heuristic. However, the above argument is invalid even for $r > 1$ and one can expect non-trivial detectability conditions on A when $r > 1$. In the dense regime we observe that simply $r > 1$ is enough for this purpose. However, the sparse regime requires a more delicate approach in terms of the effect of $r > 1$.

Remark 2.3. Note that Theorem 2.5, other than part 2(a), requires no additional assumption on θ other than the symmetry requirement in Equation (2.3).

2.6.1 Dense Regime $(\alpha \leq \frac{1}{2})$

The detection complexity in the dense regime with $r > 1$ matches the Gaussian linear model case. Interestingly, just by increasing 1 observation per treatment from the identity design matrix scenario, the detection boundary changes completely. The following theorem provides the lower and upper bound for the dense regime when $r > 1$.

Theorem 2.6. *Let* $\boldsymbol{X} \in \mathrm{SA}(r)$. *Let* $k = p^{1-\alpha}$ *with* $\alpha \leq \frac{1}{2}$ *and the block size/binomial denominator* $r > 1$.

1. *Consider the model* (2.2) *and the testing problem given by* (2.4). *Assume* $\theta \in \mathrm{BC}^1(0)$. *Then*

   (a) If $A \ll \sqrt{\frac{p^{1/2}}{kr}}$, *then all tests are asymptotically powerless.*

   (b) If $A \gg \sqrt{\frac{p^{1/2}}{kr}}$, *then the* GLRT *is asymptotically powerful.*

2. *Consider model* (2.12) *and the testing problem* (2.13). *Then*

   (a) *If* $\Delta \ll \sqrt{\frac{p^{1/2}}{kr}}$, *then all tests are asymptotically powerless.*

   (b) *If* $\Delta \gg \sqrt{\frac{p^{1/2}}{kr}}$, *then the* GLRT *is asymptotically powerful.*

Also when $\frac{A^2kr}{\sqrt{p}}$ or $\frac{\Delta^2kr}{\sqrt{p}}$ remains bounded away from 0 and ∞, the asymptotic power of GLRT remains bounded between 0 and 1. The upper and lower bound rates of the minimum signal strength match with that of Arias-Castro et al. (2011) and Ingster et al. (2010).

2.6.2 Sparse Regime $(\alpha > \frac{1}{2})$

Unlike the dense regime, the sparse regime depends more heavily on the value of r. The next theorem quantifies this result; it shows that in the sparse regime if $r \ll log(p)$, then all tests are asymptotically powerless. Indeed this can be argued from Theorem 2.1 and 2.2. However, for the sake of completeness we provide it here.

Theorem 2.7. *Let* $k = p^{1-\alpha}$ *with* $\alpha > \frac{1}{2}$. *If* $r \ll log(p)$, *then for both the problems and* (2.4) *and* (2.13), *all tests are asymptotically powerless.*

Remark 2.4. Theorem 2.7 requires no additional smoothness assumption on θ other than the symmetry requirement in Equation (2.3).

Thus, for the rest of this section we consider the case where $k \ll \sqrt{p}$ and $r \gg log(p)$. We first divide our analysis into two parts, where we study the lower bound and upper bound of the problem separately.

Lower Bound

To introduce a sharp lower bound in the regime where $\alpha > \frac{1}{2}$ and $r \gg log(p)$ in the binary regression model (2.2) and the testing problem (2.4) for the SA(r) design, we define the following functions. Figure 2.2 provides a graphical representation of the detection boundary. Define

$$\rho^*_{\text{binary}}(\alpha) = \begin{cases} \frac{(\alpha-\frac{1}{2})}{4(\theta'(0))^2} & \text{if } \frac{1}{2} < \alpha < \frac{3}{4}, \\ \frac{(1-\sqrt{1-\alpha})^2}{4(\theta'(0))^2} & \text{if } \alpha \geq \frac{3}{4}. \end{cases} \tag{2.14}$$

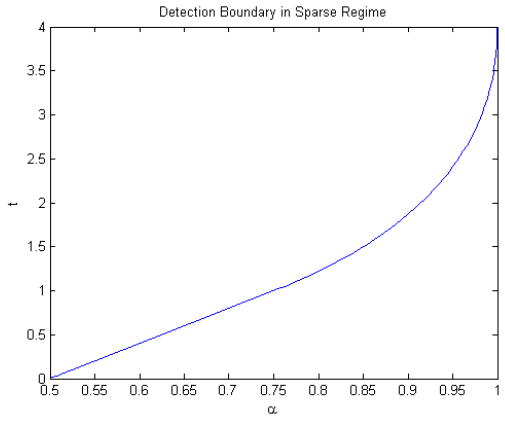

Figure 2.2: Detection boundary $t = \rho^*_{\text{binary}}(\alpha)$ in the sparse regime when θ corresponds to logistic regression. The detectable region is $t > \rho^*_{\text{binary}}(\alpha)$, and the undetectable region is $t < \rho^*_{\text{binary}}(\alpha)$. The blue curve corresponds to $t = \rho^*_{\text{binary}}(\alpha)$

This is the same as the Gaussian detection boundary (2.1) multiplied by $1/4(\theta^{'}(0))^2$. The reason for the appearance of the factor $1/4(\theta^{'}(0))^2$ is that the Fisher Information for a single Bernoulli sample under binary regression model (2.2) is equal to $\sqrt{4(\theta^{'}(0))^2}$.
For every $j \in \{1, \ldots, p\}$, we have

$$\hat{\beta}_j^{\text{MLE}} \xrightarrow{d} N(\beta_j, \sigma_j^2)$$

where $\sigma_j^2 = 4(\theta^{'}(0))^2$ under H_0 and $\sigma_j^2 \approx 4(\theta^{'}(0))^2$ under H_1. To see this, note that under H_1 we have $\sigma_j^2 = (\frac{1}{2}+\delta)(\frac{1}{2}-\delta) \approx 4\theta^{'}(0)$ where $\delta > 0$ is small and denotes a departure of the Bernoulli proportion from the null value of $\frac{1}{2}$, *i.e.*, under H_1, the outcomes corresponding to the signals follow $\text{Bernoulli}(\frac{1}{2} + \delta)$ or $\text{Bernoulli}(\frac{1}{2} - \delta)$. This implies $\sqrt{\frac{1}{4(\theta^{'}(0))^2}}\,\hat{\boldsymbol{\beta}}$ should yield a detection boundary similar to the multivariate Gaussian model case.
For the detection boundary in the corresponding binomial proportions model (2.12) and the testing problem (2.13), we define the following function

$$\rho^*_{\text{binomial}}(\alpha) = \begin{cases} \frac{(\alpha - \frac{1}{2})}{4} & \text{if } \frac{1}{2} < \alpha < \frac{3}{4} \\ \frac{(1-\sqrt{1-\alpha})^2}{4} & \text{if } \alpha \geq \frac{3}{4} \end{cases} \tag{2.15}$$

The following theorem provides the exact lower boundary for the $\text{SA}(r)$ designs for the binary regression model as well as the corresponding binomial problem.

Theorem 2.8. *Let* $\boldsymbol{X} \in \text{SA}(r)$. *Suppose* $r \gg log(p)$ *and* $k = p^{1-\alpha}$ *with* $\alpha > \frac{1}{2}$.

1. *Consider the binary regression model* (2.2) *and the testing problem* (2.4). *Further suppose that* $\theta \in \mathrm{BC}^2(0)$. *Let* $A = \sqrt{\frac{2tlog(p)}{r}}$. *If* $t < \rho^*_{\mathrm{binary}}(\alpha)$, *all tests are asymptotically powerless.*

2. *Consider the binomial model* (2.12) *and the testing problem* (2.13). *Let* $\Delta = \sqrt{\frac{2tlog(p)}{r}}$. *If* $t < \rho^*_{\mathrm{binomial}}(\alpha)$, *all tests are asymptotically powerless.*

Remark 2.5. As mentioned in the Introduction, the analysis turns out to be surprisingly nontrivial since it seems not possible to simply reduce the calculations to the Gaussian case by doing a Taylor expansion of L_π around $\boldsymbol{\beta} = 0$. In particular, a natural approach to analyze these problems is to expand the integrand of L_π by a Taylor series around $\boldsymbol{\beta} = 0$ and thereby reducing the analysis to calculations in the Gaussian situation and a subsequent analysis of the remainder term. However in order to find the sharp detection boundary, the analysis of the remainder term turns out to be very complicated and non-trivial. Thus our proof to Theorem 2.8 is not a simple application of results from the Gaussian linear model.

Upper Bound

According to Theorem 2.8, all tests are asymptotically powerless if $t < \rho^*_{\mathrm{binary}}(\alpha)$ in the sparse regime. In this section we introduce tests which reach the lower bound discussed in the previous section. We divide our analysis into two subsections. In Section $6.2.2.1$, we study the Higher Criticism Test defined by (2.10) which is asymptotically powerful as soon as $t > \rho^*_{\mathrm{binary}}(\alpha)$ in the sparse regime. In Section $6.2.2.3$, we discuss a more familiar Max Test or Minimum p-value test which attains attains the the sharp detection boundary only for $\alpha \geq \frac{3}{4}$.

The Higher Criticism Test In this section, we study the version of Higher Criticism introduced in Section 6.2. Recall, we have by Theorem 2.4 that the type I error of the Higher Criticism Test, as defined by Equation (2.11), converges to 0. The next theorem states the optimality of the Higher Criticism Test as soon as the signal strength exceeds the detection boundary.

Theorem 2.9. *Let* $\boldsymbol{X} \in \mathrm{SA}(r)$. *Suppose* $r \gg log(p)$ *and* $k = p^{1-\alpha}$ *with* $\alpha > \frac{1}{2}$.

1. *Consider the binary regression model* (2.2) *and the testing problem* (2.4). *Further suppose that* $\theta \in \mathrm{BC}^2(0)$. *Let* $A = \sqrt{\frac{2tlog(p)}{r}}$. *If* $t > \rho^*_{\mathrm{binary}}(\alpha)$ *, then the Higher Criticism Test is asymptotically powerful.*

2. *Consider the binomial model* (2.12) *and the testing problem* (2.13). *Let* $\Delta = \sqrt{\frac{2tlog(p)}{r}}$. *If* $t > \rho^*_{\mathrm{binomial}}(\alpha)$, *then the Higher Criticism Test is asymptotically powerful.*

Comparison with the Original Higher Criticism Test We begin by providing a slight simplification of T_{HC} in $\mathrm{SA}(r)$ designs. Let S be a generic $\mathrm{Bin}(r, \frac{1}{2})$ random variable and $\mathbb{B}, \overline{\mathbb{B}}$ respectively denote the distribution function and the survival function of $\frac{|S-\frac{r}{2}|}{\sqrt{\frac{r}{4}}}$. Hence

$$\mathbb{B}(t) = \mathbb{P}(\tfrac{|S-\frac{r}{2}|}{\sqrt{\frac{r}{4}}} \leq t), \overline{\mathbb{B}}(t) = 1 - \mathbb{B}(t) \ .$$

In the case of $\mathrm{SA}(r)$ designs $W_p(t) = \sqrt{p}\frac{\overline{\mathbb{F}}_p(t) - \overline{\mathbb{B}}(t)}{\sqrt{\overline{\mathbb{B}}(t)(1-\overline{\mathbb{B}}(t))}}$. The original Higher Criticism Test as defined by Donoho and Jin (2004) can also be calculated as a maximum over some appropriate function of p-values. By that token, ideally we would like to define the Higher Criticism Test statistic as

$$\mathsf{T}^{\mathrm{Ideal}}_{\mathrm{HC}} = \sup_{0<t<\frac{r}{2}} W_p(t) \ .$$

However due to difficulties in calculating the null distribution for deciding a cut-off for the rejection region, we instead work with a discretized version of it. We detail this below in the context of $\mathrm{SA}(r)$ designs. Define the j^{th} p-value as $q_j = \mathbb{P}(|\mathrm{Bin}(r, \frac{1}{2}) - \frac{r}{2}| > |Z_j - \frac{r}{2}|)$ for $1, \ldots, p$ and let $q_{(1)}, \ldots, q_{(p)}$ be the ordered p-values based on exact Binomial distribution probabilities. Define

$$\mathsf{T}'_{\mathrm{HC}} = \max_{1 \leq j \leq p} \sqrt{p}\frac{\frac{j}{p} - q_{(j)}}{\sqrt{q_{(j)}(1-q_{(j)})}} \ .$$

It is difficult to analyze the distribution of $\mathsf{T}'_{\mathrm{HC}}$ under the null to decide a valid cut-off for testing. The following proposition yields a relationship between T_{HC}, $\mathsf{T}^{\mathrm{Ideal}}_{\mathrm{HC}}$ and $\mathsf{T}'_{\mathrm{HC}}$.

Proposition 2.1. *Let* $|Z - \frac{r}{2}|_{(j)}$ *denote the* j^{th} *order statistics based on* $|Z_i - \frac{r}{2}|, \ i = 1, \ldots, p$. *For*

t such that $|Z - \frac{r}{2}|_{(p-j)} \leq t < |Z - \frac{r}{2}|_{(p-j+1)}$, *we have*

$$\sqrt{p}\frac{\overline{\mathbb{F}}_p(t) - \overline{\mathbb{B}}(t)}{\sqrt{\overline{\mathbb{B}}(t)(1-\overline{\mathbb{B}}(t))}} \leq \sqrt{p}\frac{\frac{j}{p} - q_{(j)}}{\sqrt{q_{(j)}(1-q_{(j)})}} \, .$$

Hence from Proposition 2.1, we observe that we have the following inequality

$$\mathsf{T}'_{\mathrm{HC}} \geq \mathsf{T}^{\mathrm{Ideal}}_{\mathrm{HC}} \geq \mathsf{T}_{\mathrm{HC}} \, . \tag{2.16}$$

This unlike the results in Donoho and Jin (2004) and Cai et al. (2011), where the leftmost inequality is a equality. Therefore, tt is worth further comparing the above discussion to the Higher Criticism Test introduced by (Donoho and Jin, 2004; Hall and Jin, 2010) in the Gaussian framework. In the case of orthogonal Gaussian linear models, $\mathsf{T}_{\mathrm{HC}}, \mathsf{T}'_{\mathrm{HC}}$ and $\mathsf{T}^{\mathrm{Ideal}}_{\mathrm{HC}}$ are defined by standard normal survival functions and p-values respectively and one uses Z_j instead of $\frac{Z_j - \frac{r}{2}}{\sqrt{\frac{r}{4}}}$ in the definition of T_{HC}. This yields that in the Gaussian framework the leftmost inequality of (2.16) is an equality. Moreover under the framework, standard empirical process results for continuous distribution functions yield asymptotics for $\mathsf{T}^{\mathrm{Ideal}}_{\mathrm{HC}}$ under the null. Therefore in the Gaussian case the uncountable supremum in the definition of $\mathsf{T}^{\mathrm{Ideal}}_{\mathrm{HC}}$ is attained and the statistic is algebraically equal to a maximum over finitely many functions of p-values, namely, $\mathsf{T}'_{\mathrm{HC}}$. However due to the possibility of strict inequality in Proposition 2.1 for the Binomial distribution, we cannot reduce our computation to p-values as in the Gaussian case. Although it is true that marginally each q_j is stochastically smaller than a $\mathrm{U}(0,1)$ random variable, we are unable to find a suitable upper bound for the rate of $\mathsf{T}'_{\mathrm{HC}}$ since it also depends on the joint distribution of $q_{(1)}, \ldots, q_{(p)}$. It might be possible to estimate the gaps between $\mathsf{T}'_{\mathrm{HC}}, \mathsf{T}^{\mathrm{Ideal}}_{\mathrm{HC}}$ and T_{HC}, but since this is not essential for our purpose, we do not attempt this.

Rate Optimal Upper Bound: Max Test A popular multiple comparison procedure is the minimum p-value test. In the context of Gaussian linear regression, Donoho and Jin (2004) and Arias-Castro et al. (2011) showed that the minimum p-value test reaches the sharp detection boundary if and only if $\alpha \geq \frac{3}{4}$. In this section, we introduce and study the minimum p-value test in binary regression models.

As before define the j^{th} p-value as

$$q_j = \mathbb{P}(|\text{Bin}(r, \frac{1}{2}) - \frac{r}{2}| > |Z_j - \frac{r}{2}|)$$

for $j = 1, \ldots, p$ and let $q_{(1)}, \ldots, q_{(p)}$ be the ordered p-values. We will study the test based on the minimum p-value $q_{(1)}$. Note that it is equivalent to study the test based on the statistic

$$\max_{1 \le j \le p} W_j, \quad W_j = \frac{|Z_j - \frac{r}{2}|}{\sqrt{\frac{r}{4}}} .$$

From now on, we will call this the Max Test. In the following theorem, we show that similar to Gaussian linear models, for binary regression, the Max Test attains the sharp detection boundary if and only if $\alpha \ge \frac{3}{4}$. However if one is interested in rate optimal testing, *i.e.*, only the rate or order of the detection boundary rather than the exact constants, the Max Test continues to perform well in the entire sparse regime.

Theorem 2.10. *Let $\boldsymbol{X} \in \text{SA}(r)$. Suppose $r \gg (logr)^2 log(p)$ and $k = p^{1-\alpha}$ with $\alpha > \frac{1}{2}$.*

1. *Suppose $\theta \in \text{BC}^2(0)$ and let $A = \sqrt{\frac{2tlog(p)}{r}}$. Set*

$$\rho^*_{\text{Max,binary}}(\alpha) = \frac{(1 - \sqrt{1-\alpha})^2}{4(\theta'(0))^2} .$$

Then in the model (2.2) *and problem* (2.4) *one has the following.*

   (a) If $t > \rho^*_{\text{Max,binary}}(\alpha)$, then the Max Test is asymptotically powerful.

   (b) If $t < \rho^*_{\text{Max,binary}}(\alpha)$, then the Max Test is asymptotically powerless.

2. *Let $\Delta = \sqrt{\frac{2tlog(p)}{r}}$. Set $\rho^*_{\text{Max,binomial}}(\alpha) = \frac{(1-\sqrt{1-\alpha})^2}{4}$. Then in the model* (2.12) *and problem* (2.13) *one has the following.*

   (a) If $t > \rho^*_{\text{Max,binomial}}(\alpha)$, then the Max Test is asymptotically powerful.

   (b) If $t < \rho^*_{\text{Max,binomial}}(\alpha)$, then the Max Test is asymptotically powerless.

Theorem 2.10 implies that the detection boundary for the Max Test matches the detection boundary of the Higher Criticism Test only for $\alpha \ge \frac{3}{4}$. For $\alpha < \frac{3}{4}$, the detection boundary

of the Max Test lies strictly above that of the Higher Criticism Test. Hence the Max Test fails to attain the sharp detection boundary in the moderate sparsity regime, $\alpha < \frac{3}{4}$. Thus if one is certain of high sparsity it can be reasonable to use the Max Test whereas the Higher Criticism Test performs well throughout the sparse regime. It is worth noting that the requirement $r \gg (log(r))^2 log(p)$ is a technical constraint and can be relaxed. In most situations, it does not differ much from the actual necessary condition $r \gg log(p)$ and hence we use $r \gg (log(r))^2 log(p)$ for proving Theorem 2.10.

2.7 Detection Boundary and Asymptotic Analysis for WA Designs

In this section, we study the role of the parameter vector $(n^*, n_*, r^*, r_*, Q_{n,p}, \gamma_{n,p})$ in deciding the detection boundary. For the sake of brevity, we will often drop the subscripts n, p from Q and γ when there is no confusion. Recall Ω^* from Section 2.4.

If we just concentrate on the observations corresponding to the rows with index in Ω^*, we have an orthogonal design matrix similar to $\mathrm{SA}(r)$ designs. Our proofs of lower bounds in both dense and sparse regimes and also the test statistics proposed for the attaining the sharp upper bound is motivated by this fact. Similar to $\mathrm{SA}(r)$ designs, we divide our analysis into the dense and sparse regimes. Also, owing to the possible non-orthogonality of $\boldsymbol{X}$ for WA designs, we cannot directly reduce this problem to testing homogeneity of binomial proportions as in (2.13). So, henceforth we will be analyzing model (2.2) and corresponding testing problem (2.4). However, as we shall see, under certain combinations of $(n^*, n_*, r^*, r_*, Q, \gamma)$, one can essentially treat the problem as an orthogonal design like in $\mathrm{SA}(r)$ designs. This is explained in the following two sections.

2.7.1 Dense Regime $(\alpha \leq \frac{1}{2})$

We recall the definition of the GLRT from Equation (2.9). The following theorem provides the lower and upper bound for the dense regime.

Theorem 2.11. *Let $\boldsymbol{X} \in \mathrm{WA}$. Suppose Let $k = p^{1-\alpha}$ with $\alpha > \frac{1}{2}$ and $r_* > 1$. Assume $\theta \in \mathrm{BC}^2(0)$ and set $\gamma = p^{\frac{1}{2}-\alpha}$. Then we have the following.*

1. *If* $A \ll \sqrt{\frac{p^{1/2}}{kr^*}}$, *then all tests are asymptotically powerless.*

2. *If* $A \gg \sqrt{\frac{p^{1/2}}{kr_*}}$, *then the* GLRT *is asymptotically powerful.*

We note that the form of the detection boundary is exactly same as that in Theorem 2.6 for $\mathrm{SA}(r)$ designs with r^* and r_* playing the role of r. This implies that when n_*Q^2 is not too large ($\frac{n_*Q^2}{r^*} \ll \gamma = p^{\frac{1}{2}-\alpha}$), we can still recover the same results as in $\mathrm{SA}(r)$ designs because the columns of the design matrix are weakly correlated.

2.7.2 Sparse Regime $(\alpha > \frac{1}{2})$

Unlike the dense regime, the sparse regime depends more heavily on the values of r^* and r_*. The next theorem quantifies this result; it shows that in the sparse regime if $r^* \ll log(p)$, then all tests are asymptotically powerless. This result is analogous to Theorem 2.7 for $\mathrm{SA}(r)$ designs. Indeed this can be argued from Theorem 2.1 and 2.2. However, for the sake of completeness we provide it here.

Theorem 2.12. *Let* $\boldsymbol{X} \in \mathrm{WA}$. *Let* $k = p^{1-\alpha}$ *with* $\alpha > \frac{1}{2}$ *and let* $|\bigcup_{i \notin \Omega^*} S_i| \ll p$. *If* $r^* \ll log(p)$, *then all tests are asymptotically powerless.*

Remark 2.6. The condition $|\bigcup_{i \notin \Omega^*} S_i| \ll p$, restricts the location of non-zero elements in the support of rows of $\boldsymbol{X}$ when the row has more than one non-zero element. This restriction imposes a structure on the deviation of $\boldsymbol{X}$ from orthogonality. As the proof of Theorem 2.12 will suggest, this condition ensures that the assumptions of Theorem 2.1 hold and hence renders all tests asymptotically powerless irrespective of signal strength.

The following theorem provides the value of γ that is defined in Condition C.3 in Definition 4.1, to ensure the results parallel to Theorem 2.8. Not surprisingly, the test attaining the sharp lower bound turns to be the version of the Higher Criticism Test introduced in Section 2.6. Similar to the $\mathrm{SA}(r)$ design, it is also possible to introduce and study the Max Test which attains the sharp detection boundary only for $\alpha \geq \frac{3}{4}$. However, we omit this since it can be easily derived from the existing arguments.

Theorem 2.13. *Let* $\boldsymbol{X} \in \mathrm{WA}$ *and* $k = p^{1-\alpha}$ *with* $\alpha > \frac{1}{2}$. *Suppose* $r_* \gg log(p)$, $\gamma = log(p)$, *where* γ *is defined in Definition 4.1. Further suppose that* $\theta \in \mathrm{BC}^2(0)$.

1. *Let* $A = \sqrt{\frac{2tlog(p)}{r^*}}$. *If* $t < \rho^*_{\text{binary}}(\alpha)$, *then all tests are asymptotically powerless.*

2. *Let* $A = \sqrt{\frac{2tlog(p)}{r_*}}$. *If* $t > \rho^*_{\text{binary}}(\alpha)$, *then the Higher Criticism Test is asymptotically powerful.*

Remark 2.7. The assumptions on the design matrix in Theorem 2.13 is weaker than the assumptions in Theorem 2.12. In particular, one is allowed to go beyond $|\bigcup_{i \notin \Omega^*} S_i| \ll p$ in Theorem 2.12 as long as the condition **C3** is satisfied with $\gamma = log(p)$. This is expected since the conditions under which all tests are asymptotically powerless irrespective of sample size are often more stringent.

Remark 2.8. Theorem 2.13 states that the Higher Criticism Test attains the sharp detection boundary in the sparse regime. Note that the difference in the denominators of A in the statement of upper and lower bound in Theorem 2.13 is unavoidable and the difference vanishes asymptotically if $r^*/r_* \to 1$. This is expected since the detection boundary depends on the column norms of the design matrix.

2.8 Simulation Studies

We complement our study with some numerical simulations which illustrate the empirical performance of the test statistics described in earlier sections for finite sample sizes. Since detection complexity of the general weakly correlated binary design matrices depend on the behavior of $\mathrm{SA}(r)$ type designs, we only provide simulations for strong one-way ANOVA type design. Let X be a balanced design matrix with $p = 10000$ covariates and r replicates per covariate. For different values of sparsity index $\alpha \in (0, 1)$ and r we study the performance of Higher Criticism Test, GLRT and Max Test respectively for different values of t, where t which corresponds to $A = \sqrt{\frac{2(\rho^*_{\text{logistic}}(\alpha)+t)log(p)}{r}}$.
Following (Arias-Castro et al., 2011), the performance of each of the three methods is computed in terms of the empirical risk defined as the sum of probabilities of type I and II errors achievable across all thresholds. The errors are averaged over 300 trials. Even though the theoretical calculation of null distribution of the Higher Criticism Test statistic computed from p-values remains a challenge, we performed our simulations using the p-

value based statistic $\max_{1\leq j\leq \frac{p}{2}} \sqrt{p}\frac{\frac{j}{p}-q_{(j)}}{\sqrt{q_{(j)}(1-q_{(j)})}}$ since they yielded similar expected results. Note that this statistic is different from $\mathsf{T}'_{\mathrm{HC}}$ in that the maximum is taken over the the first $\frac{p}{2}$ elements instead of all p of them. The main reason for this is the fact that, as noted by Donoho and Jin (2004), the information about the signal in the sample lies away from the extreme p-values. The GLRT is based on $\mathsf{T}_{\mathrm{GLRT}}$ as defined in Section 4.1 and the Max Test is based on the test statistic defined in Section $4.2.5$.

The results are reported in Figure 2.3 and Figure 2.4. For $r = \sqrt{log(p)} \ll log(p)$ and $k = 2, 7$ which corresponds to $k \ll \sqrt{p}$, *i.e.*, the sparse regime, we can see that all tests are asymptotically powerless in Figure 2.3 which is expected from the theoretical results. However, even when $r = \lceil\sqrt{log(p)}\rceil \ll log(p)$, for the dense regime, $k = 159$ and 631, we see from Figure 2.3 that the GLRT is asymptotically powerful whereas the other two tests are asymptotically powerless. Once r is much larger than $log(p)$ in Figure 2.4 our observations are similar to Arias-Castro et al. (2011). Here we employ simulations for $k = 2, 7, 40$ which correspond to the sparse regime and for $k = 159$ which corresponds to the dense regime. We note that the performance of GLRT improves very quickly as the sparsity decreases and begins dominating the Max Test. The performance of the Max Test follows the opposite pattern with errors of testing increasing as k increases. The Higher Criticism Test, however, continues to have good performance across the different sparsity levels once $r \gg log(p)$.

2.9 Discussions

In this paper we study testing of the global null hypothesis against sparse alternatives in the context of general binary regression. We show that, unlike Gaussian regression, the problem depends not only on signal sparsity and strength, but also heavily on a sparsity interaction parameter of the design matrix. We provide conditions on the design matrix which render all tests asymptotically powerless irrespective of signal strength. In the special case of design matrices with binary entries and certain sparsity structures, we derive the lower and upper bounds for the testing problem in both dense (rate optimal) and sparse regimes (sharp including constants). In this context, we also develop a version

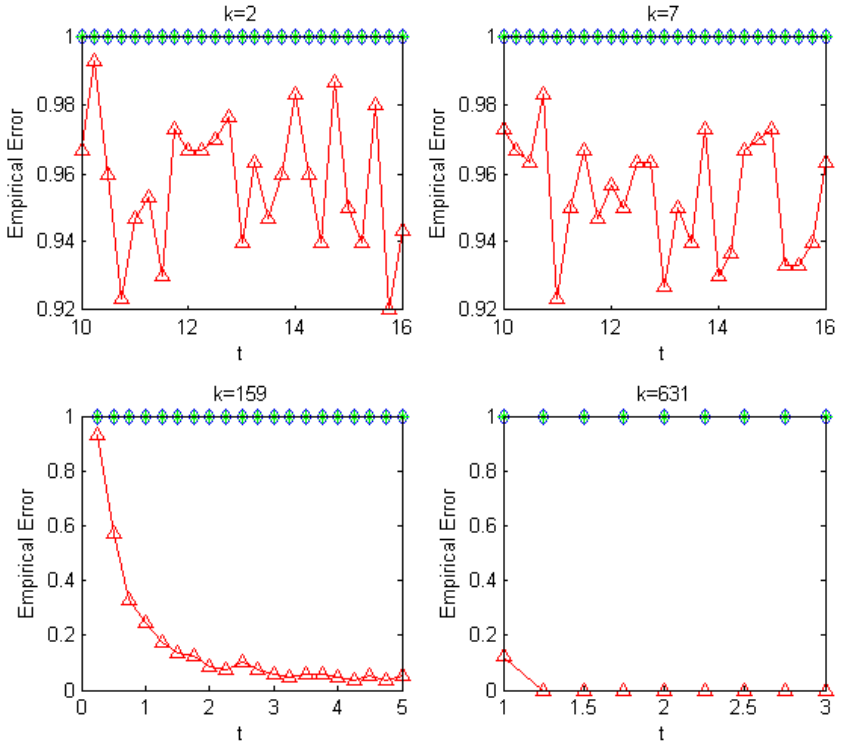

Figure 2.3: Simulation results are for $p = 10000$ and $r = \lceil \sqrt{log(p)} \rceil = 4$. Sparsity level k is indicated below each plot. In each plot, the empirical risk of each method [GLRT (red triangles); Higher Criticism (blue diamonds); Max Test (green stars)] is plotted against t which corresponds to $A = \sqrt{\frac{\max\{2(\rho^*_{\text{logistic}}(\alpha)+t),0\}log(p)}{r}}$.

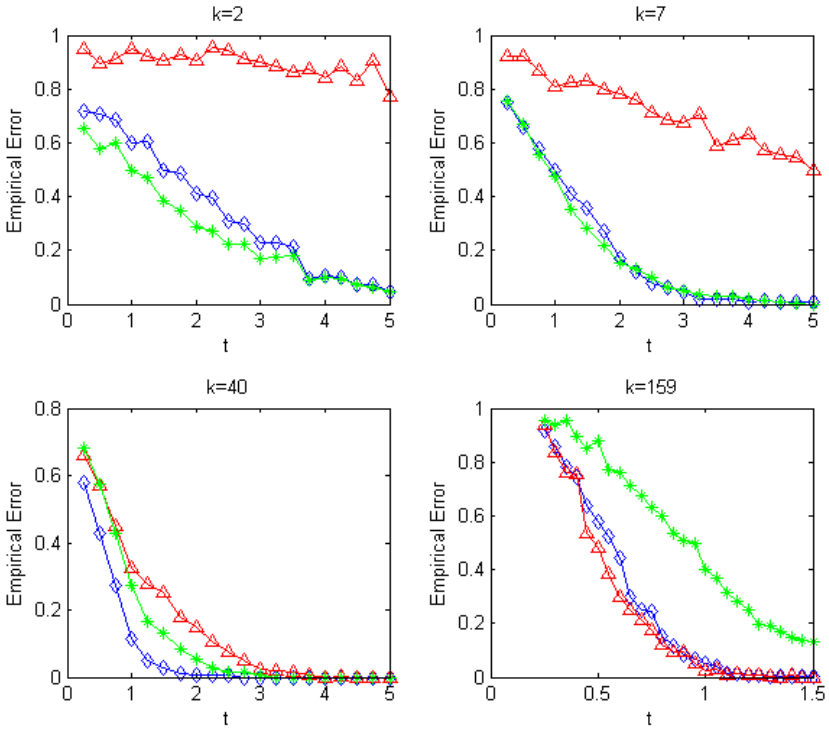

Figure 2.4: Simulation results are for $p = 10000$ and $r = \lceil (log(p))^5 \rceil = 66280$. Sparsity level k is indicated below each plot. In each plot, the empirical risk of each method [GLRT (red triangles); Higher Criticism (blue diamonds); Max Test (green stars)] is plotted against t which corresponds to $A = \sqrt{\frac{2(\rho^*_{\text{logistic}}(\alpha)+t)log(p)}{r}}$.

of the Higher Criticism Test statistic applicable for binary data which attains the sharp detection boundary in the sparse regime.

In this paper, we constructed tests by combining tests based on Z-statistics from the orthogonal part and the non-orthogonal part of the $\boldsymbol{X}$. In particular, we combine procedures based on Z_j and $Z_j^{\mathbf{G}}$ separately. This helps us achieve optimal rates for upper bounds on testing errors under the same conditions required for lower bounds in these problems. Indeed, one can consider constructing GLRT and Higher Criticism Test using Z-statistics constructed based on whole $\boldsymbol{X}$, *i.e.*, based on $Z_j^{\boldsymbol{X}} = X_j^T y,\ j = 1, \ldots, p$ directly. We could obtain similar results based on the combined Z-statistics under stronger structural assumptions on $\mathbf{G}$ than what we require here. For the sake of compactness we omit those results here and plan to study them in more detail in future research.

The study of detection boundaries associated with binary regression models for a general design matrix is much more delicate. We extend our results to allow for a more general design when the non-orthogonal columns of the design matrix are sufficiently sparse and the number of subjects with multiple non-zero entries in the design matrix are not too large. Future research is needed to extend the results to a general design matrix allowing correlation among the covariates X_j's.

Minimax Estimation in Semiparametric Regression using Higher Order Influence Functions

Rajarshi Mukherjee and James Robins

Department of Biostatistics

Harvard School of Public Health

3.1 Introduction

Robins et al. (2008) have developed a theory of point and interval estimation for non-linear functionals of the observed distribution in parametric, semi-parametric and non-parametric models based on higher order influence functions (HOIFs). HOIFs are higher order U-statistics and the theory extends the first order semi-parametric theory of Bickel et al. (1993) and Van der Vaart (1991). As derived in Robins et al. (2008), using the theory of HOIFs it is possible to produce minimax rate optimal estimators of non-linear functionals in nonparametric and semi-parametric regression problems provided the marginal density of the covariates satisfies certain lower bound on smoothness. The purpose of this paper is to understand the theory in constructing rate optimal estimators in a semi-parametric regression problem under no smoothness assumption on the marginal density of the covariates. We explain this in more detail below.

We consider the estimation of a treatment effect on an outcome in presence of a high dimensional vector $\boldsymbol{X}$ of confounding variables. Specifically, for a binary treatment A and response Y, let τ be the variance weighted average treatment effect i.e.

$$\tau := \mathbb{E}\left(\frac{Var(A|\boldsymbol{X})c(\boldsymbol{X})}{\mathbb{E}(Var(A|\boldsymbol{X})}\right) = \frac{\mathbb{E}(cov(Y,A|\boldsymbol{X}))}{\mathbb{E}(Var(A|\boldsymbol{X})} \tag{3.1}$$

where

$$c(\boldsymbol{x}) = \mathbb{E}(Y|A=1,\boldsymbol{X}=\mathbf{x}) - \mathbb{E}(Y|A=0,\boldsymbol{X}=\mathbf{x}). \tag{3.2}$$

The above follows from a simple calculation and $c(\boldsymbol{x})$ is called the average treatment effect among subjects with $\boldsymbol{X}=\boldsymbol{x}$ under the assumption of no unmeasured confounding. In this set up we are interested in the semiparametric constraint

$$c(\boldsymbol{x}) = \psi^* \text{ for all } \boldsymbol{x} \tag{3.3}$$

or specifically the model

$$\mathbb{E}(Y|A,\boldsymbol{X}) = \psi^* A + b(\boldsymbol{X}) \tag{3.4}$$

It turns out that under above model , τ equals ψ^*. Moreover, the inference on τ is closely

related to the estimation $\mathbb{E}(Cov(Y, A|\boldsymbol{X}))$ (Robins et al., 2008). Specifically, point and interval estimator for τ can be constructed from point and interval estimator of the numerator $\mathbb{E}(cov(Y, A|X))$ of τ. In particular, for any fixed $\tau^* \in \mathbb{R}$, define $Y^*(\tau^*) = Y - \tau^* A$ and the corresponding functional

$$\psi(\tau^*) = \mathbb{E}((Y^*(\tau^*) - \mathbb{E}(Y^*(\tau^*)|\boldsymbol{X}))(A - \mathbb{E}(A|\boldsymbol{X}))) = \mathbb{E}(cov(Y^*(\tau^*), A|\boldsymbol{X})).$$

Then τ is the unique solution of $\psi(\tau^*) = 0$. Suppose we can construct point estimators $\hat{\psi}(\tau^*)$ and $(1-\alpha)$ interval estimator of $\psi(\tau^*)$. Then $\hat{\tau}$ satisfying $\psi(\hat{\tau}) = 0$ is an estimator of τ with similar properties. Further a $(1-\alpha)$ confidence set for τ is the set of τ^* for which $(1-\alpha)$ interval estimator of $\psi(\tau^*)$ contains 0.

With this background, let us have n i.i.d samples of $O = (Y, A, \boldsymbol{X})$ where Y is the outcome of interest, A is a binary treatment variable and $\boldsymbol{X}$ is a set of covariates takes value in $[0,1]^d$ with typically $d \geq 10$. We are interested in

$$\mathbb{E}(Y|A, \boldsymbol{X}) = \psi^*(\theta)A + b(\boldsymbol{X}) \tag{3.5}$$

where $\psi^*(\theta)$ is an unknown parameter and $\theta = (b, p, f)$ is a vector of parameters deciding the model with $p(\boldsymbol{X}) = \mathbb{E}(A|\boldsymbol{X})$ and $f(\boldsymbol{X})$ the marginal density of $\boldsymbol{X}$. Also, we will typically denote by Θ the set of all (b, p, f) allowed in our problem. Following our above discussion, we will also discuss with estimation $\mathbb{E}(cov(Y, A|X))$ in relevant situations. When no restrictions on the treatment effect function are imposed (which we will refer to as the non-parametric model/case), *i.e.*, in the model with 3.3 unrestricted but the marginal density of $\boldsymbol{X}$ known, the minimax lower bound of estimation of τ in mean squared error norm was derived in Robins et al. (2009), which equals $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ up to multiplicative constants. Here $\overline{\beta} := \frac{\beta_b + \beta_p}{2}$ where β_b and β_p are the Hölder smoothness index of $\mathbb{E}(A|\boldsymbol{X})$ and $b(\boldsymbol{X}) := \mathbb{E}(Y|\boldsymbol{X})$ respectively. The corresponding same upper bound was obtained by Robins et al. (2008) even when f is unknown. We will see that analogous upper bound results also hold if the smoothness classes are assumed to be Sobolev balls instead Hölder. Throughout we will refer to the rate $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ as the "*proposed rate of convergence*" since unlike the nonparametric model where the minimax lower bound on rate of

convergence was derived in Robins et al. (2009), it is still unknown whether this is the optimal rate of convergence in the model defined by 3.5 . When the marginal density of $\boldsymbol{X}$ is unknown, this rate could be obtained in both semiparametric and nonparametric models under the assumption that the marginal density of $\boldsymbol{X}$ had a smoothness exponent β_f that exceeded a certain threshold that depended on $\overline{\beta}$, namely, $\frac{\beta_f}{2\beta_f+d} > \left(\frac{\beta_p \vee \beta_b}{d}\right)\left(\frac{d-4\overline{\beta}}{d+4\overline{\beta}}\right)$, where $p = \mathbb{E}(A|\boldsymbol{X})$ and $b = \mathbb{E}(Y|\boldsymbol{X})$ in the nonparametric model and $b = \mathbb{E}(Y - \psi^*(\theta)A|\boldsymbol{X})$ in the semiparametric model. Henceforth, We will often call the regime of $\overline{\beta} \geq d/4$ as regular and the regime corresponding to $\overline{\beta} < d/4$ as irregular. Under the semi-parametric model 3.5 of a constant treatment effect, one can construct an estimator that attains the proposed rate of convergence with no assumptions on the smoothness of the density of $\boldsymbol{X}$ when $\max(\beta_b, \beta_p) < 1$. However, when $\max(\beta_b, \beta_p) \geq 1$, the estimator fails to attain the proposed rate of convergence. The purpose of this paper is to understand the complexity of the problem when $\max(\beta_b, \beta_p) > 1$.

The rest of the paper is organized as follows. In Section 1 we look at the problem more formally by describing the smoothness classes under study. In Section 2, we review existing estimators producing optimal results in the nonparametric model under Hölder smoothness and provide extension of the results under Sobolev classes for univariate covariate models. We end Section 2 by reviewing a simple estimator which attains the proposed rate of convergence when $\max(\beta_b, \beta_p) < 1$ by only assuming $\beta_f > 0$. Section 3 is devoted towards developing HOIFs under a related submodel and analyzing the variance of the corresponding estimator when $\max(\beta_b, \beta_p) \geq 1$. Finally, Section 4 is devoted towards understanding the third order efficient influence function in a related submodel.

3.2 Formalizations of the Model

Model 3.5 can be stated equivalently as $\mathbb{E}\left[Y - \psi^*(\theta)A|A, \mathbf{X}\right] = b\left(\boldsymbol{X}\right)$ only depend on $\boldsymbol{X}$. Throughout we assume that the distribution of $\mathbf{X}$ is supported on the unit cube $[0,1]^d$ and has a density $f(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. It is well known that one cannot hope to have honest asymptotic confidence intervals for $\psi^*(\theta)$ shrinking in width to 0 without bounds on roughness of $b(\cdot)$ and $p(\cdot)$. For concreteness we dis-

cuss the notion of Hölder spaces as a measure of roughness of functions. Robins et al. (2008) produces minimax estimators and confidence intervals for the average treatment effect under lower bounds on smoothness of f and under both nonparametric and semi-parametric models. However, inference for $\psi^*(\theta)$ in model 3.5 without lower bounds on smoothness of f turns out to be more subtle problem. This is even true for one dimensional setting i.e. when $d = 1$. In order to study the situation of $d = 1$, we will work with a larger class of functions namely, the Sobolev spaces which we also define and discuss below. However, we first introduce some concepts of orthonormal basis and projection kernels which will be used throughout to study the finite dimensional approximations of $b(\cdot), p(\cdot)$ in appropriate function spaces. Let $\{\phi_l(\boldsymbol{x}), l = 1, 2, \ldots\}$ be a orthonormal basis (o.n.b) of $L_2(\mu)$ where μ denotes the Lebesgue measure on $[0,1]^d$. We note that by definition, $Z_l^f(\boldsymbol{x}) = \left(\mathbb{E}_f[\overline{\mathbf{Z}}_{f,k}(\boldsymbol{X})\overline{\mathbf{Z}}_{f,k}^T(\boldsymbol{X})]\right)^{-1/2}\phi_l(\boldsymbol{x}), l = 1, 2, \ldots$, with $\overline{\mathbf{Z}}_{f,k}(\boldsymbol{x}) = (Z_1^f(\boldsymbol{x}), \ldots, Z_k^f(\boldsymbol{x}))^T$, is an o.n.b of $L_2(F_{\boldsymbol{X}})$ where $F_{\boldsymbol{X}}$ stands for the marginal distribution of $\boldsymbol{X}$. Let $K_{f,k}(\boldsymbol{x}_1.\boldsymbol{x}_2) = \overline{\mathbf{Z}}_{f,k}(\boldsymbol{x}_1)^T\overline{\mathbf{Z}}_{f,k}(\boldsymbol{x}_2)$. Then for any function $h(\boldsymbol{x}) \in L_2(F_{\boldsymbol{X}})$, the projection $\Pi_f(h(\boldsymbol{x})|\overline{\mathbf{Z}}_{f,k}(\boldsymbol{x}))$ of $h(\boldsymbol{x})$ under true marginal f onto the subspace $lin\{\overline{\mathbf{Z}}_{f,k}(\boldsymbol{x})\} = \{a^T\overline{\mathbf{Z}}_{f,k}(\boldsymbol{x}) : a \in \mathbb{R}^k\}$ spanned by elements of $\overline{\mathbf{Z}}_{f,k}(\boldsymbol{x})$ is given by $\mathbb{E}_f(K_{f,k}(\boldsymbol{x}, \boldsymbol{X})h(\boldsymbol{X}))$. Thus by definition $K_{f,k}(\boldsymbol{x}, \boldsymbol{X})$ is the associated projection kernel. From now on we will denote by Π or Π_f the projection onto suitable subspaces of $L_2(F_{\boldsymbol{X}})$ under true marginal f and denote by $\hat{\Pi}$ or $\Pi_{\hat{f}}$ the orthogonal projection under estimated $\hat{f}$. By abuse of notation we will often interchangeably use both $K(\cdot,\cdot)$ and Π as projection kernels and/or projection operator. A more detailed discussion about projections onto subspaces of $L_2(\nu)$ for general measures ν on $[0,1]^d$ can be found in Appendix C. For optimal approximation in Hölder spaces one typically uses projection kernels based on compactly supported wavelet bases and for optimal approximation in Sobolev spaces for $d = 1$, one can also use projection kernel based on fourier or sine-cosine basis.

3.2.1 Hölder Spaces and Optimal Approximation

Results of Ritov and Bickel (1990) and Robins et al. (1997) imply it is not possible to construct honest asymptotic confidence intervals for $\psi^*(\theta)$ whose width shrinks to 0 as $n \to \infty$ if $b(\cdot)$ and $p(\cdot)$ are too rough. Therefore we will place the following kind of

bounds on their roughness or complexity (Robins et al., 2008).

Definition 3.1. *A function $h(\cdot)$ with domain $[0,1]^d$ is said to belong to a Hölder ball $H(\beta, C)$, with Hölder exponent $\beta > 0$ and radius $C > 0$, if and only if $h(\cdot)$ is uniformly bounded by C, all partial derivatives of $h(\cdot)$ up to order $\lfloor \beta \rfloor$ exist and are bounded, and all partial derivatives $\nabla^{\lfloor \beta \rfloor}$ of order $\lfloor \beta \rfloor$ satisfy*

$$\sup_{x, x+\delta x \in [0,1]^d} \left| \nabla^{\lfloor \beta \rfloor} h(x + \delta x) - \nabla^{\lfloor \beta \rfloor} h(x) \right| \leq C ||\delta x||^{\beta - \lfloor \beta \rfloor}.$$

It is known that the minimax rates of convergence of estimation of a marginal density or conditional expectation $h(\cdot) \in H(\beta, C)$ in $L_p, 2 < p < \infty$ and L_∞ norms are $O\left(n^{-\frac{\beta}{2\beta+d}}\right)$ and $O\left(\left(\frac{n}{log(n)}\right)^{-\frac{\beta}{2\beta+d}}\right)$ respectively. Often, estimators attaining these rates as referred to as rate optimal. Typically we will assume that $b(\cdot)$, $p(\cdot)$, and $f(\cdot)$ belong to Hölder balls $H(\beta_b, C_b)$, $H(\beta_p, C_p)$, $H(\beta_f, C_f)$. It is well known that (Härdle et al., 1998), choosing $\phi_k(\boldsymbol{x})$ to be the $\log_2 k$ level compactly supported wavelet basis with suitable vanishing moment conditions on the mother wavelet, one has

$$\sup_{h \in H(\beta, C)} \|h - \Pi_f(h|\overline{\mathbf{Z}}_{f,k})\|_2 \lesssim k^{-\beta/d}$$

.

3.2.2 Sobolev Spaces and Optimal Approximation

In order to define Sobolev spaces we first recall the Fourier or trigonometric basis of $L_2[0,1]$ defined as

$$\phi_1(x) = 1, \phi_{2k}(x) = \sqrt{2}\cos(2\pi k x), \phi_{2k+1}(x) = \sqrt{2}\sin(2\pi k x),\ k \geq 1. \tag{3.6}$$

Then for any $f \in L_2[0,1]$, let ζ_j be the Fourier coefficients of f with respect to orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ defined as

$$\zeta_j = \int_0^1 f(x)\phi_j(x)dx.$$

Also for any $\beta > 0$, define

$$a_j = \begin{cases} j^\beta & \text{if } j \text{ is even} \\ (j-1)^\beta & \text{if } j \text{ is odd} \end{cases}$$

Then we define Sobolev class of functions as follows.

Definition 3.2. *For $\beta > 0$ and $L > 0$ the Sobolev class $\tilde{W}(\beta, L)$ is defined as follows:*

$$\tilde{W}(\beta, L) = \{f \in L_2[0,1] : \zeta = (\zeta_1, \zeta_2, \ldots)^T \in \Xi(\beta, Q)\}$$

where ζ_j is the j^{th} Fourier coefficient of f and

$$\Xi(\beta, Q) = \{\zeta \in l^2(\mathbb{N}) : \sum_{j=1}^{\infty} a_j^2 \zeta_j^2 \leq Q^2\}$$

with $Q = L^2/\pi^{2\beta}$.

For all $\beta > \frac{1}{2}$ the functions belonging to $\tilde{W}(\beta, L)$ are continuous and contain the Hölder function class for integer $\beta > 0$ with same smoothness and bound L and suitable additional assumptions on the boundary value. By the definition of the Sobolev of classes it is immediate that

$$\sup_{h \in \tilde{W}(\beta, L)} \|h - \Pi_\mu(h|\overline{\mathbf{Z}}_{f,k})\|_2 \lesssim k^{-\beta}$$

where Π_μ denotes the projection kernel using Fourier basis under Lebesgue measure.

3.2.3 Assumptions and Notations

We briefly discuss the assumptions under which we work in this paper. Throughout we will assume that $\hat{b}, \hat{p}, \hat{f}$ are estimated from a separate randomly chosen training sample of size n_t satisfying $n \asymp n - n_t \asymp n$. Allowing abuse of notation, we will also denote $n_v = n - n_t$ by n since they are of the same asymptotic order. The reason for working with a split sample is because Hölder classes with $\overline{\beta} < \frac{d}{4}$ are not Donsker (Van der Vaart and Wellner, 2000). Henceforth, all expectations and variances will be interpreted as conditional on the training sample and hence will be random. However, for convenience of notation, we will often suppress this fact in the notation. Finally we will have the following same set of assumptions on the functions b, p, f as in Li et al. (2011).

(a) $|b(\boldsymbol{x})| \leq c_1$, $p(\boldsymbol{x}) \leq c_2$ with probability 1 under $F_{\boldsymbol{X}}$ for some fixed constants $0 < c_1, c_2 < \infty$.

(b) $var(Y|\boldsymbol{X}) = \sigma_Y^2(\boldsymbol{X}) < c_3$ with probability 1 under $F_{\boldsymbol{X}}$ for some fixed constants $0 < c_3 < \infty$.

(c) There exists $c_f, c_4 \in (0, \infty)$ such that $c_f < f(\boldsymbol{x}) < c_4$ for all $\boldsymbol{x}$.

(d) We suppose that $\hat{p}, \hat{b}$ are rate optimal estimators of b, p respectively with L_2 rate of convergence of order $n^{-\frac{\beta_b}{2\beta_b+d}}$ and $n^{-\frac{\beta_p}{2\beta_p+d}}$ in probability respectively.

(e) $\hat{f}$ converge to f with respect to L_p and L_∞ norm for all $p \geq 2$ at optimal rates $n^{-\frac{\beta_f}{2\beta_f+d}}$ and $(\frac{log(n)}{n})^{-\frac{\beta_f}{2\beta_f+d}}$ in probability respectively.

(f) $\hat{b}, \hat{p}, \hat{f}, 1/\hat{f}$ are uniformly bounded in supremum norm.

3.3 Preliminary Analysis

Here we first review results in minimax estimation of the average treatment effect under the nonparametric model from Li et al. (2011) where the theory is presented for Hölder balls using wavelet bases. For $d = 1$, we provide simple extensions of the theory to Sobolev ellipsoids using the fourier basis. Next in the semi-parametric regression problem 3.5, we study a simple estimator introduced by Robins et al. (2008) which attains the proposed rate of convergence under $\max(\beta_b, \beta_p) < 1$.

3.3.1 Nonparametric Model

Consider the estimation of $\psi(\theta) = \mathbb{E}_\theta(cov_\theta(Y, A|X))$ which is closely related to the estimation of $\psi^*(\theta)$ as discussed earlier. Also here we will take $b(\boldsymbol{x}) = \mathbb{E}(Y|bX)$ instead. As developed in Li et al. (2011) and Robins et al. (2008), consider the m^{th} order estimator of $\psi(\theta)$ as follows

$$\hat{\psi}_{m,k} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{b}(\boldsymbol{X}_i))(A_i - \hat{p}(\boldsymbol{X}_i)) - \sum_{j=2}^{m}\mathbb{H}_{j,j}^{(k)}$$

where

$$\mathbb{H}_{2,2}^{(k)} = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \hat{\epsilon}_{i_1} \overline{\mathbf{Z}}_{\hat{f},k}(\boldsymbol{X}_{i_1})^T \overline{\mathbf{Z}}_{\hat{f},k}(\boldsymbol{X}_{i_2}) \hat{\Delta}_{i_2}$$

and for $j \geq 3$

$$\mathbb{H}_{j,j}^{(k)} = \frac{(-1)^j}{n(n-1) \times \cdots (n-j+1)} \sum_{i_1 \neq i_2 \cdots \neq n_j} H_{j,j,\bar{i_j}}^{(k)}$$

with

$$H_{j,j,\bar{i_j}}^{(k)} = \hat{\epsilon}_{i_1} \overline{\mathbf{Z}}_{\hat{f},k}(\boldsymbol{X}_{i_1})^T \prod_{r=3}^{j} \left(\overline{\mathbf{Z}}_{\hat{f},k}(\boldsymbol{X}_{i_r}) \overline{\mathbf{Z}}_{\hat{f},k}(\boldsymbol{X}_{i_r})^T - I_{k \times k} \right) \overline{\mathbf{Z}}_{\hat{f},k}(\boldsymbol{X}_{i_2}) \hat{\Delta}_{i_2}$$

Then we have the following bias and variance properties of the estimator

Proposition 3.1. *(Li et al., 2011) Under assumptions (a)-(f) and with $\phi_l(\boldsymbol{x}), l = 1, 2, \ldots$ the tensor product elements in univariate compact wavelet basis with optimal approximation properties, for m=3,..., the estimator $\hat{\psi}_{m,k}$ has conditional bias can be decomposed into truncation bias $TB_k(\theta)$ and m^{th} order estimation bias $EB_{m,k}(\theta)$ as follows*

$$BI(\hat{\psi}_{m,k}, \psi(\theta)) = TB_k(\theta) + EB_{m,k}(\theta)$$

such that

$$TB_k = \sup_{\theta \in \Theta} TB_k(\theta) = O_p(k^{-\frac{2\overline{\beta}}{d}})$$

$$EB_m = \sup_{\theta \in \Theta} EB_{m,k}(\theta) = O_p\left(n^{-\left(\frac{\beta_p}{2\beta_p + d} + \frac{\beta_b}{2\beta_b + d} + \frac{(m-1)\beta_f}{2\beta_f + d} \right)} \right).$$

Also

$$var_\theta \asymp \frac{1}{n} \max\left(1, \left(\frac{k}{n}\right)^{m-1}\right)$$

with probability 1.

The following discussions on the implications of Proposition 3.1 is a review of results from Li et al. (2011).

Regular Case: $\overline{\beta} \geq d/4$ First assume that $\overline{\beta} \geq d/4$. Then under the assumption of $\beta_f > 0$, the estimator $\hat{\psi}_{m^*,k^*}$ with $k^* = n$ and m^* being the minimum value of m such that $\frac{\beta_p}{2\beta_p+d} + \frac{\beta_b}{2\beta_b+d} + \frac{(m-1)\beta_f}{2\beta_f+d} > \frac{1}{2}$, has mean squared error $O_p(n^{-1/2})$. Note that such a m^* always exists because $\frac{\beta_p}{2\beta_p+d} + \frac{\beta_b}{2\beta_b+d} + \frac{(m-1)\beta_f}{2\beta_f+d} > \frac{1}{2}$ is an increasing function of m when $\beta_f > 0$.

Irregular Case: $\overline{\beta} < d/4$ Now suppose $\overline{\beta} < d/4$. Then estimation of $\psi(\theta)$ at $n^{-1/2}$ is not possible. For any fixed $m \geq 2$, let $k_*(m) = n^{\frac{m}{m-1+4\overline{\beta}/d}}$ be the value of k equating the order $\frac{k^{m-1}}{n^m}$ of $var(\hat{\psi}_{m,k})$ to the order $k^{-\frac{4\overline{\beta}}{d}}$ of TB_k^2. With this choice of k, $\hat{\psi}_{m,k_*(m)}$ has optimal rate of convergence in the class $\{\hat{\psi}_{m,k} : k \in \mathbb{N}\}$ since EB_m does not depend on k. This optimal rate is given by

$$r(m) := \max\left(n^{-\left(\frac{\beta_p}{2\beta_p+d}+\frac{\beta_b}{2\beta_b+d}+\frac{(m-1)\beta_f}{2\beta_f+d}\right)}, n^{-\frac{2m\overline{\beta}/d}{m-1+4\overline{\beta}/d}}\right).$$

The optimal estimator in the class $\{\hat{\psi}_{m,k} : m \geq 2, k \geq 1\}$ is hence $\hat{\psi}_{m_*,k_*(m_*)}$ with m_* the minimizer of $r(m)$. It can be shown that if $\beta_f > d\frac{\xi(\beta_b,\beta_p,d)}{1-2\xi(\beta_b,\beta_p,d)}$ with $\xi(\beta_b,\beta_p,d) = \frac{4\overline{\beta}/d}{1+4\overline{\beta}/d} - \frac{\beta_b/d}{1+2\beta_b/d} - \frac{\beta_p/d}{1+2\beta_p/d}$ then $m_* = 2$ and hence $k_*(m_*) \asymp n^{\frac{2}{1+4\overline{\beta}/d}}$. However if $\beta_f < d\frac{\xi(\beta_b,\beta_p,d)}{1-2\xi(\beta_b,\beta_p,d)}$, $\hat{\psi}_{m_*,k_*(m_*)}$ is no longer minimax (Robins et al., 2008). However, one can construct a minimax estimator by suitably "cutting-out" certain terms from $\hat{\psi}_{m,k}$ for $m \geq 3$ and then suitably choosing $k = n^{\frac{2d}{4\overline{\beta}+d}}$ provided $\frac{\beta_f}{2\beta_f+d} > \left(\frac{\beta_p \vee \beta_b}{d}\right)\left(\frac{d-4\overline{\beta}}{d+4\overline{\beta}}\right)$.

One Dimensional Covariate Space Since expansion of functions in a suitable compactly supported wavelet basis provides optimal approximation in both Hölder and Sobolev classes, the results provided in the previous subsection extends easily for optimal mean squared error over Sobolev ellipsoids. However, it is worth noting that for $d = 1$, since fourier basis also provides optimal approximation for Sobolev spaces, it is possible to construct $\hat{\psi}_{m,k}$ using fourier basis expansion as well. The next result is for this purpose. The proof of the proposition can be found in Appendix C.

Proposition 3.2. *Suppose $d = 1$ and the classes of functions considered are Sobolev classes with the same calibration smoothness as for the Hölder classes. Also modify assumptions (a)-(f) with Sobolev classes replacing Hölder classes whenever encountered. Then with $\phi_l(\boldsymbol{x}), l = 1, 2, \ldots$ the fourier basis of $L_2[0,1]$ in equation* (3.6)*, for m=3,..., the estimator $\hat{\psi}_{m,k}$ has conditional bias and variance as in Proposition 3.1.*

Remark 3.1. *In terms of rates of convergence in different regimes of smoothness, Proposition 3.2 has the same implications as those discussed after Proposition 3.1.*

In the next section when we study the semi-parametric regression model 3.5, we shall show that even in the one dimensional case of $d = 1$, a third order estimator based on Fourier basis expansion has the incorrect order of variance. However, a third order estimator based on Haar basis expansion will be shown to have the right order of variance to attain the proposed rate of convergence of $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$.

3.3.2 Semiparametric Regression: A Simple Estimator

In this section we review a simple intuitive estimator of $\psi^*(\theta)$ developed by Robins et al. (2008) which attains the rate of convergence $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ whenever $\overline{\beta} < d/4$, $\max\{\beta_b, \beta_p\} < 1$ and $\beta_f > 0$. As a consequence, in the irregular case of $\overline{\beta} < d/4$, when the maximum smoothness of the b, p falls below unity, one can produce an estimator in the semiparametric regression problem which attains the proposed rate of convergence under no smoothness assumptions on the marginal density of the covariates. We will only assume that the unknown density $f(\boldsymbol{x})$ is absolutely continuous with respect to Lebesgue measure and both it and its inverse are bounded in sup-norm as in assumption (c).

To describe the estimator, we begin by explaining the case when $Y = A\ w.p.\ 1$ and hence the estimation of $\psi(\theta) = \mathbb{E}(var(Y|\boldsymbol{X}))$and by abuse of notation call $b(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$. Specifically suppose we divide the support of $\boldsymbol{X}$ i.e. $[0,1]^d$, into $k = k(n) = n^{\gamma}, \gamma > 1$ identical subcubes with edge length $k^{-1/d}$, where k will be decided later to suitably optimize bias and variance of the constructed estimator. It can be shown by simple probability calculation and union bound that the number of subcubes containing at least two observations is $O_p(n^2/k)$. We estimate $\mathbb{E}(var(Y|\boldsymbol{X}))$ in each such subcube by $(Y_i - Y_j)^2/2$. If for any subcube one has 3 or more observations, then i and j are chosen randomly without replacement. Let the final estimator of $\mathbb{E}(var(Y|\boldsymbol{X}))$ be the average of our subcube-specific estimates $(Y_i - Y_j)^2/2$ over the $O_p(n^2/k)$ subcubes which has at least two observations in them. Then the rate of convergence of the estimator is minimized at $n^{-\frac{4\beta/d}{4\beta/d+1}}$ by taking $k = n^{\frac{2}{1+4\beta/d}}$. In order to see this, note that $E\left[(Y_i - Y_j)^2/2|\boldsymbol{X}_i, \boldsymbol{X}_j\right] = \mathbb{E}(var(Y|\boldsymbol{X})) + \{b(\boldsymbol{X}_i) - b(\boldsymbol{X}_j)\}^2/2$, $|b(\boldsymbol{X}_i) - b(\boldsymbol{X}_j)| = O\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^{\beta_b}$ if $\beta_b < 1$ by Hölder smoothness assumption and also $\|\boldsymbol{X}_i - \boldsymbol{X}_j\| = d^{1/2}O\left(k^{-1/d}\right)$ whenever $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are in the same subcube. Therefore the estimator has variance $O_p(k/n^2)$ and

bias of order of $O\left(k^{-2\beta_b/d}\right)$. To minimize the convergence rate one can equate the orders of the variance and the squared bias by solving $k/n^2 = k^{-4\beta_b/d}$ which gives $k = n^{\frac{2}{1+4\beta_b/d}}$. However, this estimator will not converge at rate $n^{-\frac{4\beta_b/d}{4\beta_b/d+1}}$ to $E\left[var\left(Y|\boldsymbol{X}\right)\right]$ in our nonparametric model when $\beta_b > 1$, because it then no longer suffices to average estimates of $var\left(Y|\boldsymbol{X}\right)$ only over subcubes containing 2 or more observations. Now with this background we are ready to construct an estimator for $\psi^*(\theta)$ under model 3.5. The argument above implies that if $\max(\beta_b, \beta_p) < 1$, we can construct an estimator of $\hat{\tau}$ of $\psi^*(\theta)$ that converges at the rate of $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ when the semiparametric model 3.5 holds. Specifically, we again create $k = n^{\frac{2}{1+4\overline{\beta}/d}}$ subcubes. Recalling the definition $\psi(\tau)$ from the Introduction, let $\hat{\tau}$ be such that it makes the sum $\hat{\psi}(\tau)$ over subcubes containing at least 2 observations of $\{Y_i^*(\tau) - Y_j^*(\tau)\}\{A_i - A_j\}$ equal to 0 (treating subcubes with more than two observations as above), where $Y^*(\tau) = Y - \tau A$. When 3.5 holds, then $cov(Y^*(\psi^*(\theta)), A|X) = 0$. Thus an argument similar to the one above implies that $\hat{\psi}(\psi^*(\theta))$ converges to $cov(Y^*(\psi^*(\theta)), A|X) = 0$ at a rate $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$. Finally a Taylor expansion of $\hat{\psi}(\hat{\tau}) = 0$ around $\psi^*(\theta)$ yields the required rate of convergence of $\hat{\tau}$ to $\psi^*(\theta)$.

3.4 Semiparametric Regression: HOIFs under a Union Model

The previous section demonstrates an estimator which converges at the proposed rate $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ whenever $\max(\beta_b, \beta_p) < 1$ without smoothness assumption on f other than $\beta_f > 0$. Hence it is worth exploring if one can attain similar rate of convergence when $\max(\beta_b, \beta_p) \geq 1$ without further assumptions on smoothness of f. Indeed, according to results in previous sections, one can attain the proposed rate of convergence if and only if β_f exceeds a certain cut-off. However, the estimator attaining the optimal rate is based on a modified nonparametric influence function and does not use the extra information when model 3.5 holds. This section is devoted to understanding the construction of estimators which uses the extra information provided by the semi-parametric model.

It is known that higher order influence functions of $\psi^*(\theta)$ do not exist in the model 3.5 since delta dirac function is not an element of the Hilbert space L_2 of square integrable

functions (Robins et al., 2008). Hence we work under a union sub-model under which higher order influence functions of all orders exist. Since evaluating the exact higher order tangent spaces are extremely challenging, we derive a strategy to produce higher order influence functions under the union sub-model which we describe below.

Our strategy of producing higher order U-statistics estimator relies on deriving suitable HOIFs under a related submodel of 3.5. In particular, we derive HOIFs in the semiparametric regression problem under the following union submodel where either b or p has a finite expansion in terms of suitable orthonormal basis functions. We derive our influence functions under the changed assumption of normality. Therefore, unlike the previous section, the treatment in the derivation of influence function will be considered to be continuous. Our assumption of normality is for ease of computation of the HOIFs and necessary nuisance tangent spaces. At the cost of more detailed calculations it can be shown that the exact same structure of estimator is valid even for the case binary treatment. We omit such details here.

In particular, we define our working model as follows.

$$Y = \psi^*(\theta)A + b(\boldsymbol{X}) + \epsilon,\ \epsilon \sim N(0,1),$$

$$A = p(\boldsymbol{X}) + \Delta,\ \Delta \sim N(0,1).$$

To derive HOIFs we work under the union of the following two models.

Model 1

$$b(\boldsymbol{X};\eta) = \sum_{r=1}^{k} \eta_r Z_{r,f}(\boldsymbol{X}),\ \eta = (\eta_1,\ldots,\eta_k)^T \in \mathbb{R}^k,\ p(\boldsymbol{X}) \in H(\beta_p, C_p) \text{ unrestricted} \quad (3.7)$$

Model 2

$$p(\boldsymbol{X};\alpha) = \sum_{r=1}^{k} \alpha_r Z_{r,f}(\boldsymbol{X}),\ \alpha = (\alpha_1,\ldots,\alpha_k)^T \in \mathbb{R}^k,\ b(\boldsymbol{X}) \in H(\beta_b, C_b) \text{ unrestricted} \quad (3.8)$$

Throughout the following $Z_{r,i} = Z_{r,f}(\boldsymbol{X}_i)$, $\mathbf{Z}_{ki} = (Z_{1,i},\ldots,Z_{k,i})^T$ and $K(i_1,i_2) = \mathbf{Z}_{ki_1}^T \mathbf{Z}_{ki_2}$. Also the same notations with Z replaced by $\hat{Z}$ will imply changing the f to $\hat{f}$. At this point we also recall the definition of testing and estimation nuisance tangent spaces from Robins et al. (2008). Since the characterization of estimation nuisance tangent space is

more complicated, we will throughout try to understand the testing nuisance tangent space. We also note that it is easy to show that the first order efficient testing score for $\psi^*(\theta)$ in our model is given $\mathbb{ES}_1 = \sum_{i=1}^n \epsilon_i \Delta_i$. Denoting by $\Gamma_m^{nuis,test,\perp}$ the orthogonal complement of the m^{th} order testing nuisance tangent space in the space of m^{th} order zero mean U-statistics with finite variance, ideally one likes to compute the efficient testing nuisance score

$$\mathbb{ES}_m^{test} = \Pi[\mathbb{ES}_1 | \Gamma_m^{nuis,test,\perp}].$$

However, explicit computation of the above requires detailed characterization $\Gamma_m^{nuis,test,\perp}$ which in turn is a subtle problem. Hence we employ a more ad hoc procedure to evaluate possibly less efficient nuisance testing influence functions and study their asymptotic properties.

We begin by characterizing higher order testing nuisance scores in models 3.7 and 3.8. We denote the score operators by V with subscripts standing for the direction of the higher order scores and the superscript standing for the model under which the scores are derived. In particular superscripts η, α and ω stands for scores in the directions b, p and f respectively and superscript 1 or 2 stands for the scores in models 3.7 and 3.8 respectively. For the sake of concrete example $V_{\alpha\eta}^1$ stands for a second order testing nuisance score in model 3.7 in the direction α, η or b, p.

With this background, our first step in the direction of evaluating higher order testing influence functions is noticing that in model 3.7

$$V_{\alpha\eta}^1 \in \overline{lin\{\sum_{i=1}^n \epsilon_i Z_{r,i} \Delta_i g(\boldsymbol{X}_i) + \sum_{i \neq j} \epsilon_i Z_{r,i} \Delta_j g(\boldsymbol{X}_j);\ r = 1, \ldots, k, g \in L_2(F_{\boldsymbol{X}})\}}$$

which, without loss of generality, we will often write in vector form as

$$V_{\alpha\eta}^1 = \sum_{i=1}^n \epsilon_i \mathbf{Z}_{ki}^T \Delta_i g(\boldsymbol{X}_i) + \sum_{i \neq j} \epsilon_i \mathbf{Z}_{ki}^T \Delta_j g(\boldsymbol{X}_j).$$

A similar $\alpha\eta$ score in model 3.8 is

$$V_{\alpha\eta}^2 = \sum_{i=1}^n \epsilon_i \mathbf{Z}_{ki}^T \Delta_i g(\boldsymbol{X}_i) + \sum_{i \neq j} \epsilon_i \mathbf{Z}_{kj}^T \Delta_j g(\boldsymbol{X}_i).$$

An immediate consequence of the above forms of second order scores is that $V^1_{\alpha\eta}$ is the only 2nd order score not orthogonal to $\mathbb{ES}_1 = \sum_{i=1}^n \epsilon_i \Delta_i$. Further $V^1_{\alpha\eta}$ is orthogonal to all other second order scores, since clearly to not be orthogonal to $\sum_{i=1}^n \epsilon_i \mathbf{Z}^T_{ki} \Delta_j g\left(\boldsymbol{X}_j\right)$ one needs a ϵ_i and Δ_j in the score. Similarly, not to be orthogonal to $\sum_{i=1}^n \epsilon_i \mathbf{Z}^T_{ki} \Delta_i g\left(\boldsymbol{X}_i\right)$ one need both an ϵ_i and Δ_i. Analogous statement also holds in model 3.8 where the role of ϵ and Δ are simply reversed. Hence a U-statistic orthogonal to the second order testing nuisance tangent space in the union of models 3.7 and 3.8 is

$$U_{2,NP} = \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbf{Z}_{ki} \Delta_i - \frac{1}{n(n-1)} \sum_{i \neq j} \epsilon_i \mathbf{Z}^T_{ki} \mathbf{Z}_{kj} \Delta_j$$

which is the same as the second order statistic derived in the nonparametric model earlier. To see that this is indeed a testing influence function note that

$$\mathbb{E}[V^1_{\alpha\eta} U_{2,NP}] = \mathbb{E}[\mathbf{Z}^T_{ki} g(\boldsymbol{X}_i)] - \mathbb{E}[\mathbf{Z}^T_{kj} \mathbb{E}[\mathbf{Z}_{ki} \mathbf{Z}^T_{ki}] g(\boldsymbol{X}_j)] = 0$$

since $\mathbb{E}[\mathbf{Z}_{ki}\mathbf{Z}^T_{ki}] = I$ by orthonormal nature of basis functions. A similar calculation also holds for inner product with $V^2_{\alpha\eta}$. Moreover, this can be easily shown to be orthogonal to any other second order scores other than $V^1_{\alpha\eta}$ and $V^2_{\alpha\eta}$. By our arguments in previous section, the one step estimator based on $U_{2,NP}$ attains the proposed rate of convergence whenever β_f exceeds a certain cut-off. Hence in order to nullify higher order bias we proceed to derive HOIFs in our union model. The rest of the section is devoted to analyzing a third order testing influence function that demonstrates differential behavior in terms of order of asymptotic variance depending on the orthonormal basis of L_2 chosen for constructing the $\mathbf{Z}_{kj}$'s.

As before, we proceed by analyzing the third order testing nuisance scores in appropriate directions. In particular, we note that under the chosen vector notation earlier, in model 3.7

$$\begin{aligned} V^1_{\alpha\eta\omega} = &\sum_{i=1} \epsilon_i \mathbf{Z}^T_{ki} \Delta_i g(\boldsymbol{X}_i)[a(\boldsymbol{X}_i)] + \sum_{i \neq j} \epsilon_i \mathbf{Z}^T_{ki} \Delta_i g(\boldsymbol{X}_i)[a(\boldsymbol{X}_j)] \\ &+ \sum_{i \neq j} \epsilon_i \mathbf{Z}^T_{ki} \Delta_j g(\boldsymbol{X}_j)[a(\boldsymbol{X}_j)] + \sum_{i \neq j} \epsilon_i \mathbf{Z}^T_{ki}[a(\boldsymbol{X}_i)] \Delta_j g(\boldsymbol{X}_j) \end{aligned}$$

$$+ \sum_{i \neq j \neq s} \epsilon_i \mathbf{Z}_{ki}^T \Delta_j g(\boldsymbol{X}_j)[a(\boldsymbol{X}_s)]$$

where $a \in L_2(F_{\boldsymbol{X}})$ such that $\mathbb{E}(a(\boldsymbol{X})) = 0$. We have put the function a in square brackets to clarify that this arises as a score for f. A similar score $V^2_{\alpha\eta\omega}$ in model 3.8 will have the role of ϵ ad Δ reversed. Similar to before, $V^1_{\alpha\eta\omega}$ is the only 3rd order score not orthogonal to $\mathbb{ES}_1 = \sum_{i=1}^n \epsilon_i \Delta_i$ and $V^1_{\alpha\eta}$. Further, in model 3.7, $V^1_{\alpha\eta\omega}$ is itself orthogonal to all other third order scores and second order scores. Analogous statement also holds in model 3.8 where the role of ϵ and Δ are simply reversed.

Finally, going along similar lines of greedy construction of influence in the nonparametric problem discussed earlier, we define our candidate testing influence function in model 3.7 as

$$U^1_{3,candidate}(\gamma) = U_{2,NP} + \frac{1}{n(n-1)} \sum_{i \neq j} \epsilon_i \Delta_j (n-2)^{-1} \mathbf{Z}_{ki}^T \gamma(\boldsymbol{X}_i, \boldsymbol{X}_j)$$
$$+ \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq s} \epsilon_i \Delta_j \mathbf{Z}_{ki}^T \gamma(\boldsymbol{X}_i, \boldsymbol{X}_s)$$

where γ is such that $\mathbb{E}(\gamma(\boldsymbol{x}, \boldsymbol{X}_s)) = 0$ for all $\boldsymbol{x}$ and will be determined such that the $U^1_{3,candidate}(\gamma)$ is orthogonal to the third order testing nuisance tangent space. Following our previous line argument, by structure of the candidate, it suffices to be orthogonal to $V^1_{\alpha\eta\omega}$. The following proposition specifies the suitable γ for our purpose.

Proposition 3.3. *$U^1_{3,candidate}(\gamma)$ with $\gamma(\boldsymbol{X}_i, \boldsymbol{X}_s) = [I - (n-2)^{-1}\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T]^{-1}(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I)\mathbf{Z}_{ki}$ is orthogonal to $V^1_{\alpha\eta\omega}$.*

The proof of the proposition above follows by taking inner product of $U^1_{3,candidate}(\gamma)$ with each element of $V^1_{\alpha\eta\omega}$ and simple algebra shows that the inner product is 0. Hence we omit the proof here. An immediate consequence is that a similar statement holds for orthogonality with respect to $V^2_{\alpha\eta\omega}$ when $U^1_{3,candidate}(\gamma)$ has the role of subject i and subject j reversed. This immediately implies that a third order testing influence function orthogonal to the third order testing nuisance tangent space in the union of models 3.7 and 3.8 is given by

$$U_{3,IF} = A + B + C + D + E \tag{3.9}$$

where

$$A := U_{3,NP} := \frac{1}{n}\sum_i \epsilon_i \Delta_i - \frac{1}{n(n-1)}\sum_{i\neq j} \epsilon_i \mathbf{Z}_{ki}^T \mathbf{Z}_{kj} \Delta_j$$
$$+ \frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq s} \epsilon_i \Delta_j \mathbf{Z}_{kj}^T \left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I_{k\times k}\right)\mathbf{Z}_{ki}$$
$$B := \frac{1}{n(n-1)}\sum_{i\neq j} (n-2)^{-1} \epsilon_i \Delta_i \mathbf{Z}_{ki}^T \left[I_{k\times k} - \frac{(n-2)^{-1}\left\{\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T\right\}}{1+(n-2)^{-1}\left\{\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}\right\}}\right]\left(\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T - I_{k\times k}\right)\mathbf{Z}_{ki}$$
$$C := -\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq s} \epsilon_i \Delta_j \mathbf{Z}_{kj}^T \left[\frac{(n-2)^{-1}\left\{\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T\right\}}{1+(n-2)^{-1}\left\{\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}\right\}}\right]\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I_{k\times k}\right)\mathbf{Z}_{ki}$$
$$D := -\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq s} \epsilon_i \Delta_j \mathbf{Z}_{ki}^T \left[\frac{(n-2)^{-1}\left\{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\right\}}{1+(n-2)^{-1}\left\{\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}\right\}}\right]\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I_{k\times k}\right)\mathbf{Z}_{kj}$$
$$E := -\Pi[C(\boldsymbol{z}_{kj}, \boldsymbol{z}_{ks}, \mathbf{Z}_{ki})|lin\{\mathbf{Z}_{ki}\}] = -\Pi[D(\boldsymbol{Z}_{kj}, \boldsymbol{z}_{ks}, \boldsymbol{z}_{ki})|lin\{\mathbf{Z}_{kj}\}].$$

It is easy to see that by Proposition 3.3, $A+B+D$ is orthogonal to the third order testing nuisance tangent space in model 3.7 and $A+B+C$ is orthogonal to the third order testing nuisance tangent space in model 3.8. Hence by the structure of the nuisance scores, $A+B+C+D+E$ is orthogonal to the union of models 3.7 and 3.8. Once again we note that the first term corresponds to the third order influence function obtained in the nonparametric model in the previous section. In our greedy construction of HOIFs, the term B contributes the extra information present in the semiparametric model. In particular, calculation with leaving out the second order from $U_{3,candidate}(\gamma)$ corresponding to the kernel $\epsilon_i \Delta_i \mathbf{Z}_{ki}^T \gamma(\mathbf{Z}_{ki}, \mathbf{Z}_{kj})$ results in a final solution of γ which yields the final third order testing influence function in the union model to be exactly $U_{3,NP}$. Hence, the extra information from the semiparametric model can be intuitively thought to be included in the B term of the third order testing influence function. The rest of the section is devoted towards understanding the variance of $U_{3,IF}$ under Haar and Fourier basis. Finally we end the section with discussion about consequences of the respective asymptotic order of variances. The central part of the variance calculation depends on evaluating $Var(D+\tilde{A})$ where $\tilde{A} = A - U_{2,NP}$. The variance of the other terms can be shown to be of smaller order since $B + U_{2,NP}$ is a second order U-statistic and $C+E$ behaves as a remainder from an orthogonal projection. In particular, we will show that when the orthonormal basis used

is the Fourier basis, then the variance of the statistic above is not of the desired rate of k/n^2 required to achieve the proposed rate of convergence.

Variance under Haar Under the Haar basis of $L_2[0,1]^d$ orthonormal under the Lebesgue measure, it can be shown that $C + E = 0$ and also

$$\begin{aligned}
&\mathbb{E}[\epsilon_i^2\Delta_j^2(\mathbf{Z}_{ki}^T\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T\mathbf{Z}_{kj} - \frac{\mathbf{Z}_{ki}^T\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T\mathbf{Z}_{kj}}{n-2+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}})^2] \\
&= \mathbb{E}[K(i,s)K(s,j) - \frac{K(i,j)K(s,j)K(s,j)}{n-2+k}]^2 \\
&= \mathbb{E}[K(i,s)K(s,j) - \frac{kK(i,j)K(s,j)}{n-2+k}]^2 \text{ since } K(s,j)^2 = kK(s,j) \\
&= \mathbb{E}[K(i,s)K(s,j)]^2\left(\frac{n-2}{n-2+k}\right)^2 \asymp n^2 \text{ if } k \gg n
\end{aligned}$$

where the expectations are taken under the Lebesgue density. Hence, $Var(D+\tilde{A}) \asymp k/n^2$ under Haar when using true underlying distribution to construct $U_{3,IF}$ of $\boldsymbol{X}$ as uniform over $[0,1]^d$. Using this together with the assumption (f) and Lemma C.4 one can show that conditional variance of the plug in version of $D+\tilde{A}$ given the training sample is $O_p(k/n^2)$. We omit the details here.

Variance under Fourier for Univariate Co-variates In this part, we work with univariate covariates and let us first assume that they are uniformly distributed over $[0,1]$. We begin by recalling that Fourier basis of $L_2[0,1]$ is given by

$$\phi_1(x) = 1, \phi_{2k}(x) = \sqrt{2}\cos(2\pi kx), \phi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)$$

Hence, for any odd $k \geq 3$ one has $K(i,i) = k$. Also the "diagonal property", $\int K(i,s)K(s,j)d(X_s) = K(i,j)$ a.e. is satisfied for this kernel. First we show that

$$Var(D+\tilde{A}) = \left(k^2 - \left(\frac{k-2n+4}{k+n-2}\right)\frac{\mathbb{E}(K^4(j,s))}{k+n-2}\right)/n(n-1)(n-2)..$$

To see this, note that it suffices to evaluate the second moment of the kernel of the corresponding U-statistic. However, by the diagonal property

$$\mathbb{E}[\epsilon_i^2\Delta_j^2(\mathbf{Z}_{ki}^T\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T\mathbf{Z}_{kj} - \frac{\mathbf{Z}_{ki}^T\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T\mathbf{Z}_{kj}}{n-2+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}})^2]$$

$$
\begin{aligned}
&= \mathbb{E}[K^(i,s)K^2(s,j) + \frac{K^2(i,j)K^4(s,j)}{(n-2+k)^2} - 2\frac{K(i,s)K(i,j)K^3(j,s)}{n-2+k}] \\
&= \mathbb{E}[K^2(s,s)] - \left(\frac{k+2n-4}{(k+n-2)^2}\right)\mathbb{E}(K^4(j,s)) \\
&= k^2 - \left(\frac{k+2n-4}{(k+n-2)^2}\right)\mathbb{E}(K^4(j,s))
\end{aligned}
$$

as claimed. Indeed if the coefficient of k^3 in $\mathbb{E}(K^4(j,s))$ is zero then the order of the above variance is k^2/n^3. However, in the following we first show that the coefficient of k^3 is strictly positive in $\mathbb{E}K^4(i,j)$. In particular,

$$
\begin{aligned}
K^4_{2k+1}(i,j) &= (\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j) + \sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j) + 1)^4 \\
&= (\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j) + \sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))^4 + 1 \\
&+ 4(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j) + \sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))^3 \\
&+ 4(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j) + \sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j)) \\
&+ 6(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j) + \sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))^2 \\
&= (\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j))^4 + (\sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))^4 \\
&+ 4(\sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))^3(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j)) \\
&+ 4(\sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j))^3 \\
&+ 6(\sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))^2(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j))^2 + 1 \\
&+ 4(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j) + \sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))^3 \\
&+ 4(\sum_{r=1}^{k}\sqrt{2}\sin(2\pi rX_i)\sqrt{2}\sin(2\pi rX_j) + \sum_{r=1}^{k}\sqrt{2}\cos(2\pi rX_i)\sqrt{2}\cos(2\pi rX_j))
\end{aligned}
$$

$$+ 6(\sum_{r=1}^{k} \sqrt{2}\sin(2\pi r X_i)\sqrt{2}\sin(2\pi r X_j) + \sum_{r=1}^{k} \sqrt{2}\cos(2\pi r X_i)\sqrt{2}\cos(2\pi r X_j))^2$$

Hence,

$$\mathbb{E}(K_{2k+1}^4(i,j)) = \mathbb{E}(\sum_{r=1}^{k} \sqrt{2}\sin(2\pi r X_i)\sqrt{2}\sin(2\pi r X_j))^4 + R$$

where $R > 0$. To see this take for example any term in R. The proof of all other terms are similar. Consider the term: $\mathbb{E}((\sum_{r=1}^{k} \sqrt{2}\cos(2\pi r X_i)\sqrt{2}\cos(2\pi r X_j))^3(\sum_{r=1}^{k} \sqrt{2}\sin(2\pi r X_i)\sqrt{2}\sin(2\pi r X_j)))$
$= \mathbb{E}[\sum_{(r_1,r_2,r_3,r_4)\in\{1,2,3,\cdots,k\}^4} \prod_{l\in i,j} \prod_{s\in 1,2,3} \sqrt{2}\cos(2\pi r_s X_l)\sqrt{2}\sin(2\pi r_4 X_l)]$
$= \sum_{(r_1,r_2,r_3,r_4)\in\{1,2,3,\cdots,k\}^4} \mathbb{E}^2(\prod_{s\in 1,2,3} \sqrt{2}\cos(2\pi r_s X)\sqrt{2}\sin(2\pi r_4 X)) \geq 0$. The strict inequality $R > 0$ comes from the fact that one of the terms of R is 1. One more important fact is that highest power of k in any term of R is k^3. Hence if the term is non negative then the coefficient of k^3 in any of these terms must be non negative. The fact that the highest power of k in any term of R is k^3 follows along the similar lines of proof as the calculations for $\mathbb{E}(\sum_{r=1}^{k} \sqrt{2}\sin(2\pi r X_i)\sqrt{2}\sin(2\pi r X_j))^4$. Now we will just show that the coefficient of k^3 in $\mathbb{E}(\sum_{r=1}^{k} \sqrt{2}\sin(2\pi r X_i)\sqrt{2}\sin(2\pi r X_j))^4$ is positive.

Let,

$$\begin{aligned}
I_k &= \mathbb{E}(\sum_{r=1}^{k} \sqrt{2}\sin(2\pi r X_i)\sqrt{2}\sin(2\pi r X_j))^4 \\
&= I_{k-1} \\
&+ 4\mathbb{E}[\sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3} \prod_{l\in i,j} \prod_{s\in 1,2,3} \sqrt{2}\sin(2\pi r_s X_l)\sqrt{2}\sin(2\pi k X_l)] \\
&+ 6\mathbb{E}[\sum_{(r_1,r_2)\in\{1,2,3,\cdots,k-1\}^2} \prod_{l\in i,j} \prod_{s\in 1,2} \sqrt{2}\sin(2\pi r_s X_l)\sqrt{2}\sin(2\pi k X_l)\sqrt{2}\sin(2\pi k X_l)] \\
&+ 4\mathbb{E}[\sum_{r\in\{1,2,3,\cdots,k-1\}} \prod_{l\in i,j} \sqrt{2}\sin(2\pi r X_l)\sqrt{2}\sin(2\pi k X_l)\sqrt{2}\sin(2\pi k X_l)\sqrt{2}\sin(2\pi k X_l)] \\
&+ \mathbb{E}(\prod_{l\in i,j} (\sqrt{2}\sin(2\pi k X_l))^4) \\
&= I_{k-1} \\
&+ 4 \sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3} \mathbb{E}^2[\prod_{s\in 1,2,3} \sqrt{2}\sin(2\pi r_s X_l)\sqrt{2}\sin(2\pi k X_l)] \\
&+ 6 \sum_{(r_1,r_2)\in\{1,2,3,\cdots,k-1\}^2} \mathbb{E}^2[\prod_{s\in 1,2} \sqrt{2}\sin(2\pi r_s X_l)\sqrt{2}\sin(2\pi k X_l)\sqrt{2}\sin(2\pi k X_l)]
\end{aligned}$$

$$+4\sum_{r\in\{1,2,3,\cdots,k-1\}}\mathbb{E}^2[\sqrt{2}\sin(2\pi rX_l)\sqrt{2}\sin(2\pi kX_l)\sqrt{2}\sin(2\pi kX_l)\sqrt{2}\sin(2\pi kX_l)]$$
$$+\mathbb{E}^2((\sqrt{2}\sin(2\pi kX_l))^4)$$
$$\geq I_{k-1}+4\sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3}\mathbb{E}^2[\prod_{s\in 1,2,3}\sqrt{2}\sin(2\pi r_sX_l)\sqrt{2}\sin(2\pi kX_l)]$$

Now let us look at $\sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3}\mathbb{E}^2[\prod_{s\in 1,2,3}\sqrt{2}\sin(2\pi r_sX_l)\sqrt{2}\sin(2\pi kX_l)]$.

$$\begin{aligned}
&\sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3}\mathbb{E}^2[\prod_{s\in 1,2,3}\sqrt{2}\sin(2\pi r_sX_l)\sqrt{2}\sin(2\pi kX_l)]\\
=1/16&\sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3}\mathbb{E}^2(\sqrt{2}\sqrt{2}\sqrt{2}\sqrt{2}[\cos(2\pi(r_1+r_2)X_l)\cos(2\pi(k+r_3)X_l)\\
&+\cos(2\pi(r_1-r_2)X_l)\cos(2\pi(k-r_3)X_l)\\
&-\cos(2\pi(r_1-r_2)X_l)\cos(2\pi(k+r_3)X_l)\\
&-\cos(2\pi(r_1+r_2)X_l)\cos(2\pi(k-r_3)X_l)])\\
=4/16&\sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3}\mathbb{E}^2([\sqrt{2}\cos(2\pi(r_1+r_2)X_l)\sqrt{2}\cos(2\pi(k+r_3)X_l)\\
&+\sqrt{2}\cos(2\pi(r_1-r_2)X_l)\sqrt{2}\cos(2\pi(k-r_3)X_l)\\
&-\sqrt{2}\cos(2\pi(r_1-r_2)X_l)\sqrt{2}\cos(2\pi(k+r_3)X_l)\\
&-\sqrt{2}\cos(2\pi(r_1+r_2)X_l)\sqrt{2}\cos(2\pi(k-r_3)X_l)])
\end{aligned}$$

Now for each tuple $(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3$ only one of the four terms of the terms survive in the sum $\mathbb{E}[\sqrt{2}\cos(2\pi(r_1+r_2)X_l)\sqrt{2}\cos(2\pi(k+r_3)X_l)+\sqrt{2}\cos(2\pi(r_1-r_2)X_l)\sqrt{2}\cos(2\pi(k-r_3)X_l)-\sqrt{2}\cos(2\pi(r_1-r_2)X_l)\sqrt{2}\cos(2\pi(k+r_3)X_l)-\sqrt{2}\cos(2\pi(r_1+r_2)X_l)\sqrt{2}\cos(2\pi(k-r_3)X_l)]$. This combined with the fact that we are taking square of expectation implies that the negative signs will contribute an absolute value in the sum. Also, each time a term survives the expectation it equals 1 in value. This is because we are using an orthonormal basis. Now we can count the number of terms each contribute. The number of times $\sqrt{2}\cos(2\pi(r_1+r_2)X_l)\sqrt{2}\cos(2\pi(k+r_3)X_l)$ have non-zero expectation is the number of tuples $(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3$ such that $r_1+r_2=k+r_3$. The number of such terms is given by $\frac{(k-1)(k-2)}{2}$ and for each contribution to $\mathbb{E}^2$ is 1. The number of

terms for which $\sqrt{2}\cos(2\pi(r_1-r_2)X_l)\sqrt{2}\cos(2\pi(k-r_3)X_l)$ have non-zero expectation is the number of tuples $(r_1, r_2, r_3) \in \{1, 2, 3, \cdots, k-1\}^3$ such that $|r_1 - r_2| = k - r_3$. The number of such terms is given by $2\frac{(k-1)(k-2)}{2}$ and for each contribution to $\mathbb{E}^2$ is $1.\sqrt{2}\cos(2\pi(r_1 - r_2)X_l)\sqrt{2}\cos(2\pi(k + r_3)X_l)$ never survive in expectation. The number of terms for which $\sqrt{2}\cos(2\pi(r_1 + r_2)X_l)\sqrt{2}\cos(2\pi(k - r_3)X_l)$ have non-zero expectation is the number of tuples $(r_1, r_2, r_3) \in \{1, 2, 3, \cdots, k-1\}^3$ such that $r_1 + r_2 + r_3 = k$. The number of such terms is given by $\frac{(k-1)(k-2)}{2}$ and for each contribution to $\mathbb{E}^2$ is 1. Hence
$\sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3} \mathbb{E}^2[\prod_{s\in 1,2,3} \sqrt{2}\sin(2\pi r_s X_l)\sqrt{2}\sin(2\pi k X_l)] = 2(k-1)(k-2)$.
Hence,

$$\begin{aligned} I_k &\geq I_{k-1} + 4/16 \sum_{(r_1,r_2,r_3)\in\{1,2,3,\cdots,k-1\}^3} \mathbb{E}^2[\prod_{s\in 1,2,3} \sqrt{2}\sin(2\pi r_s X_l)\sqrt{2}\sin(2\pi k X_l)] \\ &= I_{k-1} + 2(k-1)(k-2)/4 \end{aligned}$$

Hence adding up from 1 to k we get that the coefficient of k^3 is positive. Now by a similar argument as above, it can be shown that $\mathbb{E}(K^4(s,j))$ is indeed a polynomial in k with degree 3. Finally since the expression for $Var(D + \tilde{A})$ is valid for any n, the coefficient of k^3 cannot be strictly greater than 1. By more careful combinatorial calculations as above, one can also calculate the exact coefficient of the polynomial and show that the coefficient of k^3 is strictly less than 1. In particular, closer look at the calculation above shows that the only terms that can possibly contribute to the coefficient of k^3 in the binomial expansion of $K^4(s,j)$ are the fourth power terms in the binomial expansion. This can be argued from looking at how k^3 arises in the final polynomial form of $\mathbb{E}(K^4(s,j))$. The k^3 term arises from a recursive argument which in turn depends on the number of positive integer solutions of an equation with less than or equal to 3 variables with absolute value of each solution less than or equal to $k-1$. The equations in question contribute a k^3 term in the final solution of the recursion if and only if the equation has exactly 3 unknowns. However, by our recursion scheme this happens if and only if we are solving the recursion for the fourth power term. For the sake of compactness we do not show such extensive calculations here. Finally, this implies that the order of variance of $U_{3,IF}$ asymptotically behaves like $\frac{k^2}{n^3}$ under Fourier basis orthonormal under the Lebesgue measure. Since we

are interested in minimax rates of estimation, this implies that worst case variance for $U_{3,IF}$ based on Fourier basis orthonormal under general f is also at least of the order $\frac{k^2}{n^3}$. Below we discuss the consequences of this asymptotic order of variance.

Consequences An immediate consequence of the variance of $U_{3,IF}$ being of the order k^2/n^3 is that by matching the order of variance with the truncation bias of the estimator, as discussed in previous sections, does not yield the proposed rate of convergence as $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$. However, since under the Haar basis, the order of the variance continues to be k/n^2 one can still attain the $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ of convergence provided β_f exceeds some deterministic cut-off. However, since Haar only provides required optimal order of approximation in Hölder spaces when $\max(\beta_b, \beta_p) < 1$, one cannot reach the optimal order of truncation bias using Haar wavelet to construct the testing influence function when $\max(\beta_b, \beta_p) \geq 1$. However when $\max(\beta_b, \beta_p) < 1$, we have already discussed a much simpler estimator in the previous section which attains $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ rate of convergence. On the other hand when $\max(\beta_b, \beta_p) \geq 1$, the use of Fourier basis to construct $U_{3,IF}$ yields incorrect order of variance. Since the construction of HOIFs are greedy in nature, the order of variance remains a problem even in higher orders. Indeed, it remains a question whether it is possible to cleverly cut-out certain terms to reduce variance without exploding the bias. Similar ideas in the nonparametric influence functions typically exploits certain smoothness assumptions on the marginal density f. However, it remains an open problem to derive the minimal requirement on the smoothness of f required for successful attainment of $n^{-\frac{4\overline{\beta}}{4\overline{\beta}+d}}$ of convergence in the semiparametric regression problem.

3.5 Towards Third Order Efficient Influence Function

In the previous section we observed that there exists a third order testing influence function in the union of the models 3.7 and 3.8 which attains the desired order of variance when using Haar wavelets but fails to attain the proposed rate of variance under Fourier basis. There are two immediate implications of this. The first one is that the inefficiency in the variance terms arises from not using the efficient influence function. The second one

is that there does not exist any third order testing influence function with the desired rate of variance other than under Haar wavelets. In particular, it suffices to understand if the variance behaves like or exceeds k/n^2 in asymptotic order. In this token, in this section we try to characterize the third order efficient score in model 3.7. By symmetry this also yields characterization of the efficient score in model 3.8. Indeed if the efficient testing influence function in the smaller model does not have the desired order of variance then indeed it is not possible to attain the correct order of variance by any other testing influence function. However, if the variance of the efficient testing influence function attains the desired order of variance, one needs to derive the efficient testing influence function in the union of models 3.7 and 3.8 to proceed further. This is because the truncation bias $TB_k(\theta)$ in any of these two models alone fails to be of the correct order and only in the union of the two models does the truncation bias simplifies as a product of two tails obtained by truncating both b and p upto finitely many basis functions. Thus, as a first essential step, it is important to characterize the efficient testing score in models 3.7 and 3.8.

Towards the above mentioned goal recall that

$$\mathbb{ES}_m^{test} = \Pi[\mathbb{ES}_1 | \Gamma_m^{nuis,test,\perp}].$$

In particular, since $V_{\alpha\eta}^1$ is the only 2nd order score in model 3.7 not orthogonal to $\mathbb{ES}_1 = \sum_{i=1}^n \epsilon_i \Delta_i$ and $V_{\alpha\eta}^1$ is orthogonal to all other second order, we have

$$\mathbb{ES}_2 = \Pi[\mathbb{ES}_1 | lin\{V_{\alpha\eta}^1\}^\perp]$$

where by abuse of notation we denote by $\mathbb{ES}_2$ as the second order efficient testing score in model 3.7. By a similar argument

$$\mathbb{ES}_3 = \Pi[\mathbb{ES}_1 | lin\{V_{\alpha\eta}^1, V_{\alpha\eta\omega}^1\}^\perp]$$

Therefore in order to evaluate the efficient scores it is necessary to first characterize $lin\{V_{\alpha\eta}^1\}^\perp$ and $lin\{V_{\alpha\eta}^1, V_{\alpha\eta\omega}^1\}^\perp$. Towards this end, we recall that in the vector form notation, the second order score operator $V_{\alpha\eta}^1$ can be indexed by a function $g \in L_2(F_{\boldsymbol{X}})$ as

follows

$$V_{\alpha\eta}^{1}(g) = \sum_{i} \epsilon_i \mathbf{Z}_{ki}^{T} g(\boldsymbol{X}_i) \Delta_i + \sum_{i \neq j} \epsilon_i \mathbf{Z}_{ki}^{T} \Delta_j g(\boldsymbol{X}_j)$$

Here and in the following, by abuse of notation, we write a function $h(\boldsymbol{X})$ taking values in $\mathbb{R}^q$ for some $q \geq 1$ as $h \in L_2(F_{\boldsymbol{X}})$ if $\mathbb{E}(h^T h) < \infty$ under the Lebesgue measure and hence also under general f by assumption (f). Hence $lin\{V_{\alpha\eta}^{1}\}$ can be described by

$$\mathcal{V}_{\alpha\eta}^{1} = \left\{ V_{\alpha\eta}^{1}(g) \, ; g \in L_2(F_{\boldsymbol{X}}) \text{ unrestricted} \right\}$$

Similarly, we index the third order score by two functions r, a as follows:

$$V_{\alpha\eta\omega}^{1}(r,a) = \begin{array}{c} \sum_i \epsilon_i \mathbf{Z}_{ki}^{T} r(\boldsymbol{X}_i) \Delta_i a(\boldsymbol{X}_i) + \sum_{i \neq j} \epsilon_i \mathbf{Z}_{ki}^{T} \Delta_j r(\boldsymbol{X}_j) \{a(\boldsymbol{X}_i) + a(\boldsymbol{X}_j)\} \\ + \sum_{i \neq j} \epsilon_i \mathbf{Z}_{ki}^{T} \Delta_i r(\boldsymbol{X}_i) [a(\boldsymbol{X}_j)] + \sum_{i \neq j \neq s} \epsilon_i \mathbf{Z}_{ki}^{T} \Delta_j r(\boldsymbol{X}_j) [a(\boldsymbol{X}_s)] \end{array}$$

with $r, a \in L_2(F_{\boldsymbol{X}})$ such that $\mathbb{E}[a(\boldsymbol{X})] = 0$. Hence Hence the $lin\{V_{\alpha\eta\omega}^{1}\}$ can be described by

$$\mathcal{V}_{\alpha\eta\omega}^{1} = \left\{ V_{\alpha\eta\omega}^{1}(r,a) \, ; r, a \in L_2(F_{\boldsymbol{X}}), \ \mathbb{E}[a(\boldsymbol{X})] = 0 \right\}$$

The next proposition provides exact form of the second order efficient score in model 3.7. The proof can be found in Appendix C.

Proposition 3.4. *The second order efficient testing score in model 3.7 is given by*

$$\begin{aligned} \mathbb{ES}_2 &= \mathbb{ES}_1 - \Pi\left[\mathbb{ES}_1 | \mathcal{V}_{\alpha\eta}^{1}\right] \\ &= \frac{1}{n} \sum_{i} \epsilon_i \left\{ \frac{n(n-1)}{n-1+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}} \right\} \Delta_i - \frac{1}{n(n-1)} \sum_{i \neq j} \epsilon_i \mathbf{Z}_{ki}^{T} \mathbf{Z}_{kj} \frac{n(n-1)}{n-1+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}} \Delta_j \end{aligned}$$

By Proposition 3.4, note that if $\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki} = k$, as is the case for Haar or Fourier basis, then

$$\mathbb{ES}_2 = \left\{ \frac{n(n-1)}{n-1+k} \right\} \left\{ \frac{1}{n} \sum_{i} \epsilon_i \Delta_i - \frac{1}{n(n-1)} \sum_{i \neq j} \epsilon_i \Delta_j \mathbf{Z}_{ki}^{T} \mathbf{Z}_{kj} \right\}$$

which is similar to $U_{2,NP}$ up to a multiple of $n(n-1)/(n-1+k)$. Therefore the bias and variance properties of the second order efficient testing influence function in model 3.7 is similar to $U_{2,NP}$ and discussions in Section 3.4 and Section 3.2.3 apply. Therefore we

proceed to evaluate the third order efficient testing score in model 3.7 as follows:

$$\begin{aligned}\mathbb{ES}_3 &= \mathbb{ES}_1 - \Pi\left[\mathbb{ES}_1 | lin\left\{\mathcal{V}^1_{\alpha\eta}, \mathcal{V}^1_{\alpha\eta\omega}\right\}\right] \\ &= \mathbb{ES}_2 - \Pi\left[\mathbb{ES}_1 | lin\left\{\mathcal{V}^1_{\alpha\eta\omega} - \Pi\left[\mathcal{V}^1_{\alpha\eta\omega}|\mathcal{V}^1_{\alpha\eta}\right]\right\}\right]\end{aligned}$$

where

$$\mathcal{V}^1_{\alpha\eta w} - \Pi\left[\mathcal{V}^1_{\alpha\eta w}|\mathcal{V}^1_{\alpha\eta}\right] = \{V^1_{\alpha,\eta,\omega}(r,a) - \Pi\left[V^1_{\alpha,\eta,\omega}(r,a)|\mathcal{V}^1_{\alpha\eta}\right]; r, a \in L_2(F_{\boldsymbol{X}}),\ \mathbb{E}\left[a\left(\boldsymbol{X}\right)\right] = 0\}.$$

Hence in order to characterize $\mathbb{ES}_3$ we will first analyze the space $lin\left\{\mathcal{V}^1_{\alpha\eta w} - \Pi\left[\mathcal{V}^1_{\alpha\eta w}|\mathcal{V}^1_{\alpha\eta}\right]\right\}$ followed by $\Pi\left[\mathbb{ES}_1|\left\{\mathcal{V}^1_{\alpha\eta w} - \Pi\left[\mathcal{V}^1_{\alpha\eta w}|\mathcal{V}^1_{\alpha\eta}\right]\right\}\right]$.

Characterizing $lin\left\{\mathcal{V}^1_{\alpha\eta w} - \Pi\left[\mathcal{V}^1_{\alpha\eta w}|\mathcal{V}^1_{\alpha\eta}\right]\right\}$

We begin by recalling that any third order mixed $\alpha\eta\omega$ score can be written as

$$\begin{aligned}V^1_{\alpha\eta\omega}\left(\tilde{r}, a\right) = \frac{1}{n}\sum_i \frac{1}{(n-2)(n-1)}\epsilon_i \mathbf{Z}_{ki}^T r\left(\boldsymbol{X}_i\right)\Delta_i a\left(\boldsymbol{X}_i\right) \\ + \left\{(n-1)\,n\right\}^{-1}\sum_{i\neq j}\frac{1}{(n-2)}\epsilon_i \mathbf{Z}_{ki}^T \Delta_j r\left(\boldsymbol{X}_j\right)\left\{a\left(\boldsymbol{X}_i\right) + a\left(\boldsymbol{X}_j\right)\right\} \\ + \left\{(n-1)\,n\right\}^{-1}\sum_{i\neq j}\frac{1}{(n-2)}\epsilon_i \mathbf{Z}_{ki}^T \Delta_i r\left(\boldsymbol{X}_i\right)\left[a\left(\boldsymbol{X}_j\right)\right] \\ + \left\{(n-2)(n-1)\,n\right\}^{-1}\sum_{i\neq j\neq s}\epsilon_i \mathbf{Z}_{ki}^T \Delta_j r\left(\boldsymbol{X}_j\right)\left[a\left(\boldsymbol{X}_s\right)\right]\end{aligned}$$

where $\tilde{r} = \frac{r}{n(n-1)(n-2)}$ and the rescaling of r is for convenience of algebraic manipulation. Now, for any (r,a) we have $\Pi\left[V^1_{\alpha\eta}\left(\tilde{r},a\right)|\mathcal{V}^1_{\alpha\eta}\right] = V_{\alpha,\eta}(g^*)$ where we find the g^* using the definition of projection as follows

$$\mathbb{E}\left[V^1_{\alpha\eta\omega}\left(\tilde{r},a\right)V^1_{\alpha\eta}\left(g\right)\right] = \mathbb{E}\left[V^1_{\alpha\eta}\left(g\right)V^1_{\alpha\eta}\left(g^*\right)\right]$$

for all $g \in L_2(F_{\boldsymbol{X}})$. Thus

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{(n-2)(n-1)} a(\boldsymbol{X}_i)\, r(\boldsymbol{X}_i)^T \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T g(\boldsymbol{X}_i)\right] \\
&+\mathbb{E}\left[\frac{1}{(n-2)} r(\boldsymbol{X}_j)^T \mathbb{E}\left[a(\boldsymbol{X}_i)\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T\right] g(\boldsymbol{X}_j)\right] \\
&+\mathbb{E}\left[\frac{1}{(n-2)} r(\boldsymbol{X}_j)^T g(\boldsymbol{X}_j)\, a(\boldsymbol{X}_j)\right] \\
&=\mathbb{E}\left[\frac{1}{(n-1)} g^*(\boldsymbol{X}_i)^T \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T g(\boldsymbol{X}_i)\right] + \mathbb{E}\left[g^*(\boldsymbol{X}_i)^T g(\boldsymbol{X}_j)\right]
\end{aligned}$$

Since the above holds for all $g \in L_2(F_{\boldsymbol{X}})$, one has

$$\begin{aligned}
&\frac{1}{(n-2)(n-1)} a(\boldsymbol{X}_i)\, r(\boldsymbol{X}_i)^T \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T \\
&+\frac{1}{(n-2)} r(\boldsymbol{X}_i)^T \mathbb{E}\left[a(\boldsymbol{X}_i)\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T\right] + \frac{1}{(n-2)} a(\boldsymbol{X}_i)\, r(\boldsymbol{X}_i)^T \\
= &\frac{1}{(n-1)} g^*(\boldsymbol{X}_i)^T \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T + g^*(\boldsymbol{X}_i)^T
\end{aligned}$$

Hence

$$\begin{aligned}
\left\{I + \frac{1}{(n-1)}\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T\right\} g^*(\boldsymbol{X}_i) = &\frac{1}{(n-2)(n-1)} \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T r(\boldsymbol{X}_i)\, a(\boldsymbol{X}_i) \\
&+ \frac{1}{(n-2)}\mathbb{E}\left[a(\boldsymbol{X})\mathbf{Z}_k\mathbf{Z}_k^T\right] r(\boldsymbol{X}_i) + \frac{1}{(n-2)} a(\boldsymbol{X}_i)\, r(\boldsymbol{X}_i)
\end{aligned}$$

Solving for g^* one has,

$$\mathbf{Z}_{ki}^T g^*(\boldsymbol{X}_i) = \mathbf{Z}_{ki}^T \left\{\frac{a(\boldsymbol{X}_i)}{(n-2)} + \frac{\mathbb{E}\left[a(\boldsymbol{X})\mathbf{Z}_k\mathbf{Z}_k^T\right]}{(n-2)}\left(\frac{(n-1)}{(n-1)+\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}}\right)\right\} r(\boldsymbol{X}_i)$$

Similarly,

$$\begin{aligned}
\mathbf{Z}_{ki}^T g^*(\boldsymbol{X}_j) = &\mathbf{Z}_{ki}^T a(\boldsymbol{X}_j)\, r(\boldsymbol{X}_j)\frac{1}{(n-2)} \\
&+ \mathbf{Z}_{ki}^T \left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}\frac{1}{(n-2)}\mathbb{E}\left[a(\boldsymbol{X})\mathbf{Z}_k\mathbf{Z}_k^T\right] r(\boldsymbol{X}_j)
\end{aligned}$$

Therefore, finally we have a characterization of the projection space as follows

$$
\begin{aligned}
& \Pi\left[V_{\alpha\eta\omega}^{1}\left(r,a\right)|\mathcal{V}_{\alpha\eta}^{1}\right] \\
= & \frac{1}{n\left(n-1\right)}\sum_{i}\mathbf{Z}_{ki}^{T}\left\{\begin{array}{c} a\left(\boldsymbol{X}_{i}\right)\frac{1}{(n-2)} \\ +\frac{1}{(n-2)}\left(\frac{(n-1)}{(n-1)+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right)\mathbb{E}\left[a\left(\boldsymbol{X}\right)\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}\right]\end{array}\right\}r\left(\boldsymbol{X}_{i}\right)\epsilon_{i}\Delta_{i} \\
& +\left\{\left(n-1\right)n\right\}^{-1}\sum_{i\neq j}\epsilon_{i}\Delta_{j}\left[\begin{array}{c} \mathbf{Z}_{ki}^{T}a\left(\boldsymbol{X}_{j}\right)r\left(\boldsymbol{X}_{j}\right)\frac{1}{(n-2)} \\ +\mathbf{Z}_{ki}^{T}\left\{I-\frac{\mathbf{z}_{kj}\mathbf{z}_{kj}^{T}}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right\}\frac{1}{(n-2)}\mathbb{E}\left[a\left(\boldsymbol{X}\right)\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}\right]r\left(\boldsymbol{X}_{j}\right)\end{array}\right] \\
= & \frac{1}{n}\sum_{i}\mathbf{Z}_{ki}^{T}\left\{\begin{array}{c} a\left(\boldsymbol{X}_{i}\right)\frac{1}{(n-1)(n-2)} \\ +\frac{1}{(n-2)}\left(\frac{1}{(n-1)+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right)\mathbb{E}\left[\left(a\left(\boldsymbol{X}\right)\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right)\right]\end{array}\right\}r\left(\boldsymbol{X}_{i}\right)\epsilon_{i}\Delta_{i} \\
& +\left\{\left(n-1\right)n\right\}^{-1}\sum_{i\neq j}\epsilon_{i}\Delta_{j}\left[\begin{array}{c} \left\{\mathbf{Z}_{ki}^{T}a\left(\boldsymbol{X}_{j}\right)r\left(\boldsymbol{X}_{j}\right)\frac{1}{(n-2)}\right\} \\ +\mathbf{Z}_{ki}^{T}\left\{I-\frac{\mathbf{z}_{kj}\mathbf{z}_{kj}^{T}}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right\}\frac{\mathbb{E}\left[a(\boldsymbol{X})\mathbf{z}_{k}\mathbf{z}_{k}^{T}\right]}{(n-2)}r\left(\boldsymbol{X}_{j}\right)\end{array}\right]
\end{aligned}
$$

Since the functions r, a always appear as a product in the above expression, by denoting $\gamma(\boldsymbol{X}_1, \boldsymbol{X}_2) = a(\boldsymbol{X}_2)r(\boldsymbol{X}_1)$ we have the residual projection space indexed by γ as

$$
\begin{aligned}
V_{\alpha\eta\omega}^{1:2}\left(\gamma\right) := & V_{\alpha\eta\omega}^{1}\left(\gamma\right)-\Pi\left[V_{\alpha\eta\omega}^{1}\left(\gamma\right)|\mathcal{V}_{\alpha\eta}^{1}\right] \\
& -\frac{1}{n}\textstyle\sum_{i}\mathbf{Z}_{ki}^{T}\left\{\frac{1}{(n-2)}\left(\frac{1}{(n-1)+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right)\mathbb{E}\left[\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}\gamma\left(\boldsymbol{X}_{i},\boldsymbol{X}\right)\right]\right\}\epsilon_{i}\Delta_{i} \\
& +\left\{\left(n-1\right)n\right\}^{-1}\textstyle\sum_{i\neq j}\frac{1}{(n-2)}\epsilon_{i}\mathbf{Z}_{ki}^{T}\Delta_{j}\gamma\left(\boldsymbol{X}_{j},\boldsymbol{X}_{i}\right) \\
= & -\left\{\left(n-1\right)n\right\}^{-1}\textstyle\sum_{i\neq j}\epsilon_{i}\Delta_{j}\left\{\begin{array}{c} \frac{1}{(n-2)}\mathbf{Z}_{ki}^{T} \\ \left\{I-\frac{\mathbf{z}_{kj}\mathbf{z}_{kj}^{T}}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right\}\mathbb{E}\left[\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}\gamma\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right]\end{array}\right\} \\
& +\left\{\left(n-1\right)n\right\}^{-1}\textstyle\sum_{i\neq j}\frac{1}{(n-2)}\epsilon_{i}\mathbf{Z}_{ki}^{T}\Delta_{i}\gamma\left(\boldsymbol{X}_{i},\boldsymbol{X}_{j}\right) \\
& +\left\{\left(n-2\right)\left(n-1\right)n\right\}^{-1}\textstyle\sum_{i\neq j\neq s}\epsilon_{i}\mathbf{Z}_{ki}^{T}\Delta_{j}\gamma\left(\boldsymbol{X}_{j},\boldsymbol{X}_{s}\right)
\end{aligned}
$$

Now we are ready to proceed towards $\chi(\gamma^*) = \Pi\left[\mathbb{ES}_1|\left\{\mathcal{V}_{\alpha\eta w}^{1}-\Pi\left[\mathcal{V}_{\alpha\eta w}^{1}|\mathcal{V}_{\alpha\eta}^{1}\right]\right\}\right]$ for the appropriate $\gamma^*(\boldsymbol{X}_1, \boldsymbol{X}_2) = a^*(\boldsymbol{X}_1)r^*(\boldsymbol{X}_2)$ deciding the projection in question. In particular, the leading term in the variance of estimator based on $\mathbb{ES}_3 = \mathbb{ES}_2 - \Pi\left[\mathbb{ES}_1|\left\{\mathcal{V}_{\alpha\eta w}^{1}-\Pi\left[\mathcal{V}_{\alpha\eta w}^{1}|\mathcal{V}_{\alpha\eta}^{1}\right]\right\}\right]$ is:

$$
\begin{aligned}
& var\left[\left\{\left(n-2\right)\left(n-1\right)n\right\}^{-1}\sum_{i\neq j\neq s}\epsilon_{i}\Delta_{j}\mathbf{Z}_{ki}^{T}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}_{s}\right)\right] \\
\leq & \frac{1}{n^{3}}\mathbb{E}\left[\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}_{s}\right)^{T}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}_{s}\right)\right]
\end{aligned}
$$

Therefore we only need to determine the optimal $\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}_s\right)$ and calculate the second moment for $k = n^{\frac{2}{1+4\beta/d}}$. Towards this end we note that the γ^* is in general deter-

mined by $\Pi[\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq s}\epsilon_i\Delta_j\mathbf{Z}_{ki}^T(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T-I)\mathbf{Z}_{kj}|lin\{V_{\alpha\eta\omega}^{1:2}(\gamma)\}]$ - the third order part of $A=U_{3,NP}$ derived Section 3. The following proposition provides an expression for this optimal γ^*. However, we will first require the following notations. To begin with , we note that the following set of calculations is provided assuming $\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}=\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}=k\ w.p.1$. The proof in Appendix C can however be traced to provide the result in terms of general forms of $\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}$ and $\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}$ and we omit the details here. In particular, this assumption is inspired by the fact that for the Haar basis orthonormal under Lebesgue measure $\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}=\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}=k$ and for the Fourier basis also orthonormal under Lebesgue measure one has $\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}=\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}=\frac{k-1}{2}$ when k is odd. With this in mind, let $c_n=\frac{1}{n-1}\left(\frac{n-1}{n-1+k}\right)^2$, $d_n=\frac{1}{n-1}$ and

$$\mathbf{A}_{sj}=I+\frac{\mathbf{A}_j+\mathbf{A}_s}{n-2}$$

with $\mathbf{A}_s=\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T$, $B_s=A_s-I$ and $\mathbf{A}_j=\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T$. Also let

$$\mathbf{D}=\mathbf{I}-\frac{\mathbf{A}_j}{n-1+k}$$

and by denoting $\mathbf{G}_{sj}=-\mathbf{A}_s^T(n-2+k)-\mathbf{A}_j^T(n-2+k)+(\mathbf{Z}_{ks}\mathbf{Z}_{kj}^T+\mathbf{Z}_{kj}\mathbf{Z}_{ks}^T)K(s,j)$ we have

$$\mathbf{A}_{sj}^{-1}=\mathbf{I}+\frac{\mathbf{G}_{sj}}{(n-2+k)^2-K^2(s,j)}.$$

Finally suppose

$$\mathbf{W}=\mathbf{M}^{-1}\mathbf{v}$$

with $\mathbf{M}=\mathbb{E}_s\left[d_n\mathbf{A}_s\mathbf{A}_{sj}^{-1}\mathbf{B}_s\left(c_n\mathbf{A}_j-2\mathbf{D}-\mathbf{B}_s^{-1}+\mathbf{D}^2+\mathbf{I}\right)\right]$ and $\mathbf{v}=\mathbb{E}_s\left[\mathbf{A}_s\mathbf{A}_{sj}^{-1}\mathbf{B}_s\mathbf{Z}_{kj}\right]$ where $\mathbb{E}_s$ denotes expectation with respect to $\boldsymbol{X}_s$.

Proposition 3.5. *With the above notations, one has*

$$\gamma^*(\boldsymbol{X}_s,\boldsymbol{X}_j)=\mathbf{A}_{sj}^{-1}\left[\mathbf{B}_s\mathbf{Z}_{kj}-d_n\{c_n\mathbf{B}_s\mathbf{A}_j-2\mathbf{B}_s\mathbf{D}-\mathbf{I}+\mathbf{B}_s\mathbf{D}^2\}\right]\mathbf{W} \tag{3.10}$$

Remark 3.2. *Although Proposition 3.5 provides a formula for the optimal γ^* which decides the variance of the third order efficient testing score in the model 3.7, the calculation of the order of variance turns out to be extremely difficult. In particular, usual techniques of upper bounding the*

variance yields sub-optimal rates of the variance of k^2/n^3. Further the calculation of the efficient testing score in the union of the models 3.7 and 3.8 requires another projection of the efficient third order score evaluated in model 3.7 onto the orthocomplement of the third order testing nuisance tangent space in the model 3.8. Deriving reasonably explicit formula for this projection also turns out to be quite challenging and remains a open problem.

3.6 Discussions

In this paper, we have studied estimation of average treatment effect of a treatment on an outcome in a semiparametric regression model using the theory of higher order influence functions. In particular, we were interested in the situation where the covariates are random and there are no smoothness assumptions on the marginal density of the covariates. We observe surprising dependence on the orthonormal basis of $L_2(\mathbb{R}^d)$ chosen for construction of the estimators where d is the dimension of the covariate space. To be more specific, the mean squared error of a particular third order testing influence function is different from when one uses a Haar basis to when one uses a Fourier basis. Although we do not explicitly derive the exact order of variance for general compactly supported wavelet in this paper, we believe that similar phenomenon continue to hold for general wavelets as well. This is a previously unheard phenomenon and raises interesting questions in the context of minimax behavior of the problem. We also characterize the third order efficient influence function in a submodel for this problem which might be useful for future research. Our future research is targeted towards understanding the variance of the third order efficient influence function in the union of models 3.7 and 3.8. In particular, if one is able to show that the variance of the efficient influence function does not attain the correct order variance for Fourier or wavelet bases, then either $n^{-\frac{4\bar{\beta}}{4\bar{\beta}+d}}$ cannot be achieved by our method of HOIFs or more interestingly the structure of the minimax rate of convergence changes depending on whether $\max(\beta_b, \beta_p) < 1$ or $\max(\beta_b, \beta_p) \geq 1$. This is a previously unheard phenomenon and requires further research for better understanding.

References

ARIAS-CASTRO, E., CANDÈS, E. J., PLAN, Y. ET AL. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics* **39** 2533–2556.

BAI, Z. D. and SARANADASA, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6** 311–329.

BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606.

BICKEL, P. J., KLAASSEN, C. A., BICKEL, P. J., RITOV, Y., KLAASSEN, J., WELLNER, J. A. and RITOV, Y. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

CAI, T. T., JENG, J. X. and JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 629–662.

CARTER, A. and POLLARD, D. (2004). Tusnádys inequality revisited. *The Annals of Statistics* **32** 2731–2741.

CHEN, S. X. and QIN, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38** 808–835.

DICKER, H. B., L. and LIN, X. (2012). Variable selection and estimation with the seamless-l0 penalty. *Statist. Sinica* .

DONOHO, D. L. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on* **57** 5467–5484.

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961.

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics* 1686–1732.

HÄRDLE, W., KERKYACHARIAN, G., TSYBAKOV, A. and PICARD, D. (1998). *Wavelets, approximation, and statistical applications*. Springer.

INGSTER, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics* **6** 47–69.

INGSTER, Y. I. (1998). Minimax detection of a signal for lˆ n-balls. *Mathematical Methods of Statistics* **7** 401–428.

INGSTER, Y. I. and SUSLINA, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*, vol. 169. Springer.

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics* **4** 1476–1526.

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **32** 111–131.

LEE, S., ABECASIS, G., BOEHNKE, M. and LIN, X. (2014). Analysis of Rare Variants in Sequencing-based Association Studies. *The American Journal of Human Genetics, under revision* .

LI, L., TCHETGEN TCHETGEN, E., VAN DER VAART, A. and ROBINS, J. M. (2011). Higher order inference on a treatment effect under low regularity conditions. *Statistics & probability letters* **81** 821–828.

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Annals of Statistics* **40** 1637.

LOPES, J. L. J., M. E. and WAINWRIGHT, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *arXiv preprint arXiv:1108.2401* .

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37** 3498–3528.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34** 1436–1462.

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* 246–270.

PAN, Z. and ZHANG, C. (2013). Relax sparse eigenvalue conditions for sparse estimation via non-convex regularized regression. *arXiv preprint arXiv:1306.3343* .

PLAN, Y. and VERSHYNIN, R. (2013a). One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics* **66** 1275–1297.

PLAN, Y. and VERSHYNIN, R. (2013b). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on* **59** 482–494.

RITOV, Y. and BICKEL, P. J. (1990). Achieving information bounds in non and semiparametric models. *The Annals of Statistics* 925–938.

ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. W. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman, ed. by D. Nolan, and T. Speed. Beachwood, OH: Institute of Mathematical Statistics* 335–421.

ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. W. (2013). Higher Order Estimating Equations .

ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. W. (2009). Semiparametric minimax rates. *Electronic Journal of Statistics* **3** 1305–1321.

ROBINS, J. M., RITOV, Y. ET AL. (1997). Toward a curse of dimensionality appropriate(coda) asymptotic theory for semi-parametric models. *Statistics in medicine* **16** 285–319.

TANG, H., JIN, X., LI, Y., JIANG, H., TANG, X., YANG, X., CHENG, H., QIU, Y., CHEN, G., MEI, J. ET AL. (2014). A large-scale screen for coding variants predisposing to psoriasis. *Nature genetics* **46** 40–50.

VAN DER VAART, A. W. (1991). On differentiable functionals. *The Annals of Statistics* 178–204.

VAN DER VAART, A. W. (2000). *Asymptotic statistics*, vol. 3. Cambridge university press.

VAN DER VAART, A. W. and WELLNER, J. A. (2000). *Weak Convergence and Empirical Processes*. Springer.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

VICTOR, R. G., HALEY, R. W., WILLETT, D. L., PESHOCK, R. M., VAETH, P. C., LEONARD, D., BASIT, M., COOPER, R. S., IANNACCHIONE, V. G., VISSCHER, W. A.

ET AL. (2004). The dallas heart study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *The American journal of cardiology* **93** 1473–1480.

WALD, A. (1950). *Statistical decision functions.* Chelsea Publishing Co.

WILLIAMS, D. (1991). *Probability with martingales*. Cambridge University Press.

ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.

ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **27** 576–593.

ZHAO, P. and YU, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7** 2541–2563.

Appendix A

Proofs for Chapter 1

Technical Lemmas

Gaussian Tail Bounds

Lemma A.1. *For $t > 1$*

$$(1 - \frac{1}{t^2})\frac{\phi(t)}{t} \leq \overline{\Phi}(t) \leq \frac{\phi(t)}{t}$$

Proof. The proof of this fact is standard and can be found in (Williams, 1991). □

Sub-Gaussian Design Matrices

Lemma A.2. *Suppose $\boldsymbol{Z}_{n\times m}$ is sub-Gaussian with parameters (Σ, K). Let the columns of $\boldsymbol{Z}$ be $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$. Then for any $\epsilon \in (0, 1)$ one has*

$$\mathbb{P}\{\max_{1\leq j\leq m} \|\boldsymbol{z}_j\|^2 > (1+\epsilon)n\} \leq 2me^{-\frac{M\epsilon^2}{K^2}n},$$

where $M > 0$ is a constant.

Proof. Denoting the elements of $\boldsymbol{Z}$ by $(Z_{ij})_{n\times m}$, note that $\sup_{i,j} \|Z_{ij}\|_{\psi^2} \leq K$ since for any i, j, $\|Z_{ij}\|_{\psi^2} \leq \|\langle e_i^T \boldsymbol{Z}, e_j\rangle\|_{\psi^2} \leq K$ by assumption. Hence by Vershynin (2010), we have that for all $j = 1, \ldots, p$,

$$\mathbb{P}(|\sum_{i=1}^{n} Z_{ij}^2 - \sum_{i=1}^{n} E(Z_{ij}^2)| \geq \delta n) \leq 2e^{-M\min(\frac{\delta^2}{4K^4}, \frac{\delta}{2K^2})n}$$

for some constant $M > 0$. Since $Z_{i,j}$'s are centered and scaled, this implies that $K \geq 1/\sqrt{2}$

and $E(Z_{ij}^2) = 1$ for all i, j. Therefore, $\delta/2K^2 \leq 1$ whenever $\delta \leq 1$. Hence we have for any $1 > \epsilon > 0$,

$$\begin{aligned}\mathbb{P}(|\sum_{i=1}^{n} Z_{ij}^2| > (1+\epsilon)n) &\leq \mathbb{P}(|\sum_{i=1}^{n} Z_{ij}^2 - \sum_{i=1}^{n} E(Z_{ij}^2)| \geq (1+\epsilon)n - \sum_{i=1}^{n} E(Z_{ij}^2)) \\ &= \mathbb{P}(|\sum_{i=1}^{n} Z_{ij}^2 - \sum_{i=1}^{n} E(Z_{ij}^2)| \geq \epsilon n) \leq 2e^{-M\frac{\epsilon^2}{4K^4}n}.\end{aligned}$$

So finally by union bound, we have

$$\mathbb{P}(\max_{1\leq j\leq m} \|\boldsymbol{z}_j\|^2 > (1+\epsilon)n) \leq 2me^{-M\frac{\epsilon^2}{4K^4}n}.$$

□

Lemma A.3. *Suppose $\boldsymbol{Z}$ is sub-Gaussian with parameters (Σ, K). Let the sample covariance matrix be $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z}$. If $log(p) \ll n$, there exists a constant c_0 such that*

$$\mathbb{P}\left(\max_{i,j} |\widehat{\boldsymbol{\Sigma}}_{ij} - \boldsymbol{\Sigma}_{i,j}| > c_0 K^2 \sqrt{\frac{log(p)}{n}}\right) \leq c_1 e^{-c_2 log(p)}$$

for positive constants c_1, c_2.

Proof. This can be proved using Lemma 14 in the supplement of Loh and Wainwright (2012). □

Lemma A.4. *Let $\boldsymbol{Z}$ be a $n \times m$ random matrix whose rows are centered i.i.d sub-Gaussian random vectors with covariance matrix $\boldsymbol{\Sigma}$. Let the sample covariance matrix be $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z}$. Then with probability at least $1 - 2e^{-ct^2}$, we have*

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \leq \max\{\delta, \delta^2\} \textit{where } \delta = C\sqrt{\frac{m}{n}} + \frac{t}{\sqrt{n}}.$$

Here $C = C_K, c = c_K > 0$ depend only on the sub-Gaussian norm $K = \|Z_1\|_{\psi^2}$ and $\|\cdot\|_2$ denotes the spectral norm a square matrix.

Proof. This can be proved using Vershynin (2010). □

Properties of $\widehat{\boldsymbol{\beta}}_\lambda$

We first begin by a brief review of oracle property of concave penalized least squares estimators based on Zhang (2010). For $\boldsymbol{\beta} \in \mathbb{R}^p$, recall $O(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$. Let $\boldsymbol{\Omega}_{O(\boldsymbol{\beta})} = \frac{\boldsymbol{X}^T_{O(\boldsymbol{\beta})}\boldsymbol{X}_{O(\boldsymbol{\beta})}}{n}$. Given the knowledge of $O(\boldsymbol{\beta})$, the oracle least square estimator $\widehat{\boldsymbol{\beta}}^o = (\hat{\beta}^o_1, \ldots, \hat{\beta}^o_p)^T$ is given by

$$(\hat{\beta}^o_j, j \in O(\boldsymbol{\beta}))^T = \frac{\boldsymbol{\Omega}^{-1}_{O(\boldsymbol{\beta})}\boldsymbol{X}^T_{O(\boldsymbol{\beta})}y}{n},\ (\hat{\beta}^o_j, j \notin O(\boldsymbol{\beta}))^T = 0$$

provided $\boldsymbol{X}_{O(\boldsymbol{\beta})}$ is of full column rank. Let

$$(w^o_j, j \in O(\boldsymbol{\beta})) = \text{ the diagonal elements of } \boldsymbol{\Omega}^{-1}_{O(\boldsymbol{\beta})}$$

so that $\mathrm{Var}(\hat{\beta}^o_j|\boldsymbol{X}) = \frac{w^o_j}{n}$ for $j \in O(\boldsymbol{\beta})$. We will also say that *global convexity criterion* holds if

$$s_{\min}(\boldsymbol{X}^T\boldsymbol{X}/n) + \frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1} > 0,\ \forall 0 < t_1 < t_2.$$

For any $v \in \mathbb{R}^q$ such that $\sum_{j=1}^q I(v_j \neq 0) = q$, define

$$\kappa(\rho; v) = \lim_{\epsilon \to 0+} \max_{1 \leq j \leq q} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1}.$$

By concavity of ρ in condition **(C1)**, $\kappa(\rho; v) \geq 0$ and by mean value theorem, it is easy to show that $\kappa(\rho; v) = \max_{1 \leq j \leq q} -\rho''(|v_j|)$ provided the second derivative of ρ is continuous. For SCAD penalty, $\kappa(\rho; v) = 0$ unless some component $|v| := (|v_1|, \ldots, |v_q|)$ takes value in $[\lambda, a\lambda]$ in which case the value is $\lambda^{-1}(a-1)^{-1}$. For the MCP penalty, a similar thing happens, i.e., $\kappa(\rho; v) = 0$ unless some component of $|v|$ takes value in $[0, a\lambda]$ in which case it equals to a^{-1}. The following lemma is about oracle variable selection property of concave penalized estimators.

Lemma A.5. *(a) Let $\lambda > 0$ be fixed and $\widehat{\boldsymbol{\beta}}_\lambda$ the penalized likelihood estimator of $\boldsymbol{\beta}$ with penalty p_λ satisfying* **(C1′)** *with constant c. Let $\hat{O} = \{j : \beta_{\lambda,j} \neq 0\}$. Suppose $s_{\min}(\Omega_{O(\boldsymbol{\beta})})$ is bounded away*

above 0 *and that* $p \leq n$. *If* $\min\{|\beta_j| : \beta_j \neq 0\} \geq c\lambda$, *then*

$$\mathbb{P}(\hat{O} \neq O(\boldsymbol{\beta})|\boldsymbol{X}) \leq \mathbb{P}(\widehat{\boldsymbol{\beta}}^o \neq \widehat{\boldsymbol{\beta}}_\lambda|\boldsymbol{X}) \leq \sum_{j \in O(\boldsymbol{\beta})} \Phi\left(\frac{c\lambda - |\beta_j|}{\sqrt{w_j^o/n}}\right) + 2\sum_{j \notin O(\boldsymbol{\beta})} \Phi\left(-\frac{n\lambda\rho^{'}(0+)}{\|\boldsymbol{x}_j\|}\right).$$

(b) Suppose $\boldsymbol{X}$ *is sub-Gaussian with parameters* (Σ, H) *such that* $\frac{1}{s_{\min}(\Sigma)} = O(1)$ *and* $s_{\max}(\Sigma) \ll p^\epsilon$ *for all* $\epsilon > 0$. *Then if* $p \leq n$ *and penalty* p_λ *satisfies* **(C1′)** *with some constant* c

$$\mathbb{P}(\hat{O} \neq O(\boldsymbol{\beta})) \leq \mathbb{P}(\widehat{\boldsymbol{\beta}}^o \neq \widehat{\boldsymbol{\beta}}_\lambda) \to 0$$

as $p \to 0$ *uniformly in* $A \gg \sqrt{\frac{log(p)}{n}}$ *provided* $\lambda = \frac{1}{\rho^{'}(0+)}\sqrt{\frac{2(1+\epsilon_p)log(p)}{n}}$ *for some* $\epsilon_p > 0$ *which can be allowed to converge to* 0 *at a suitable rate.*

Proof. (a) Following the arguments of Fan and Lv (2011), $\widehat{\boldsymbol{\beta}}_\lambda \in \mathbb{R}^p$ is a local minimizer of $Q_n(\boldsymbol{\beta})$ if the following happens

$$\begin{aligned} \boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda)/n = \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{\lambda,j})p_\lambda^{'}(|\widehat{\boldsymbol{\beta}}_{\lambda,j}|) \text{ for } \widehat{\boldsymbol{\beta}}_{\lambda,j} \neq 0 \\ |\boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\lambda)/n| < \lambda\rho^{'}(0+) \text{ for } \widehat{\boldsymbol{\beta}}_{\lambda,j} = 0 \\ s_{\min}(\boldsymbol{X}_O^T\boldsymbol{X}_O/n) > \lambda\kappa(\rho; \widehat{\boldsymbol{\beta}}_{\lambda,O}), \end{aligned}$$

where $O = O(\widehat{\boldsymbol{\beta}}_\lambda)$ is the support of $\widehat{\boldsymbol{\beta}}_\lambda$. On the other hand, if $\widehat{\boldsymbol{\beta}}_\lambda$ is a local minimizer of $Q_n(\boldsymbol{\beta})$, then the above three conditions hold with strict inequalities replaced by non-strict inequalities. Following the arguments of (Zhang, 2010), we thus have that, $\widehat{\boldsymbol{\beta}}^o$ is a solution of the above conditions in the event

$$\begin{aligned} U(\lambda) := \{\min_{j \in O(\boldsymbol{\beta})} \operatorname{sgn}(\beta_j)\widehat{\beta}_j^o > c\lambda\} \cap \{\max_{j \notin O(\boldsymbol{\beta})} |\boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n| < \lambda\rho^{'}(0+)\} \\ \cap \{s_{\min}(\Omega_{O(\boldsymbol{\beta})}) > \lambda\kappa(\rho; \widehat{\boldsymbol{\beta}}^o_{O(\boldsymbol{\beta})})\} \end{aligned}$$

since the penalty function satisfies **(C1′)**. Also by **(C1′)** and definition of κ,

$$\{\min_{j \in O(\boldsymbol{\beta})} \operatorname{sgn}(\beta_j)\widehat{\beta}_j^o > c\lambda\} \cap \{s_{\min}(\Omega_{O(\boldsymbol{\beta})}) > \lambda\kappa(\rho; \widehat{\boldsymbol{\beta}}^o_{O(\boldsymbol{\beta})})\} = \{\min_{j \in O(\boldsymbol{\beta})} \operatorname{sgn}(\beta_j)\widehat{\beta}_j^o > c\lambda\}.$$

if $s_{\min}(\Omega_{O(\boldsymbol{\beta})})$ is bounded away above 0. Now note that given $\boldsymbol{X}$, $\widehat{\boldsymbol{\beta}}^o \sim \mathrm{N}(\beta_j, w_j^o/n)$. Since

$|\beta_j| > c\lambda$ for $j \in O(\boldsymbol{\beta})$, we have for such j's

$$\mathbb{P}(\text{sgn}(\beta_j)\widehat{\beta}_j^o > c\lambda) \leq \Phi\left(\frac{c\lambda - |\beta_j|}{(w_j^o/n)^{1/2}}\right).$$

Using (Zhang, 2010), $\boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n$ are mean 0 normal random variables with variance

$$\text{Var}(\boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n) \leq \frac{\|\boldsymbol{x}_j\|^2}{n^2}.$$

Hence

$$\mathbb{P}(|\boldsymbol{x}_j(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^o)/n| \geq \lambda\rho^{'}(0+)) \leq 2\Phi\left(-\frac{n\lambda\rho^{'}(0+)}{\|\boldsymbol{x}_j\|}\right)$$

for $j \notin O(\boldsymbol{\beta})$. The desired result then follows by union bound.

(b) In order to prove the result note that it suffices to bound the following with high probability uniformly in $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying $\boldsymbol{\beta} \in \Theta_k^A$ with $A \gg \sqrt{\frac{log(p)}{n}}$:

$$\sum_{j \in O(\boldsymbol{\beta})} \overline{\Phi}\left(\frac{c\lambda - \|\beta_j|}{\sqrt{w_j^0/n}}\right) + 2\sum_{j \notin O(\boldsymbol{\beta})} \overline{\Phi}\left(-\frac{n\lambda\rho^{'}(0+)}{\|\boldsymbol{X}_j\|}\right)$$

and also guarantee that uniformly in $\boldsymbol{\beta} \in \Theta_k^A$ one has that $\frac{1}{s_{\min}\left(\Omega_{O(\boldsymbol{\beta})}\right)} = O_p(1)$. The proof of the latter fact will be included in proving the bound on the first part. We explain this more below. Now by Lemma A.4

$$w_j^0 = e_j^T\left(\frac{\boldsymbol{X}_{O(\boldsymbol{\beta})}^T\boldsymbol{X}_{O(\boldsymbol{\beta})}}{n}\right)e_j \leq \frac{1}{\left(s_{\min}\left(\frac{\boldsymbol{X}_{O(\boldsymbol{\beta})}^T\boldsymbol{X}_{O(\boldsymbol{\beta})}}{n}\right)\right)}$$

$$\leq \frac{1}{s_{\min}(\Sigma) - \xi_p(n,p,k,H)}$$

with probability at least $1 - e^{-c_H k}$. Here $\xi_p(n,p,k,H) = C_H\sqrt{\frac{k}{n}}$ and one can take $c_H = O(1/H^4)$ and $C_H = O(1)$ (Vershynin, 2010). Also by assumption $\max(s_{\max}(\Sigma), H) \ll p^{\epsilon}$ for all $\epsilon > 0$ hence $e^{-c_H k} = o(1)$. Hence with probability $1 - o(1)$, uniformly in $\boldsymbol{\beta} \in \Theta_k^A$ such that $A \gg \sqrt{\frac{log(p)}{n}}$, one has the following for some $\delta_p \to 0$ whenever $\lambda = \sqrt{\frac{2log(p)(1+\epsilon_p)}{(\rho^{'}(0+))^2 n}}$ with

$\epsilon_p = O(1)$:

$$\sum_{j \in O(\boldsymbol{\beta})} \overline{\Phi}\left(\frac{c\lambda - |\beta_j|}{\sqrt{w_j^0/n}}\right) \leq \sum_{j \in O(\boldsymbol{\beta})} \overline{\Phi}\left(-\sqrt{log(p)}\delta_p(s_{\min}(\Sigma) - \xi_p(n,p,k,H))\right)$$
$$\leq \frac{k}{p} e^{-\kappa_p}.$$

The last inequality above follows from Lemma A.1 with some $\kappa_p \to \infty$ since $\frac{1}{s_{\min}(\Sigma)} = O(1)$ and $\xi_p(n,p,k,H) = o(1)$ by assumptions of the lemma. This completes the desired bound on the first of the two terms. Note that, under the assumptions of the lemma, $\frac{1}{s_{\min}(\Sigma)} = O(1)$ and $\xi_p(n,p,k,H) = o(1)$ also implies that uniformly in $\boldsymbol{\beta} \in \Theta_k^A$ one has that $\frac{1}{s_{\min}\left(\Omega_{O(\boldsymbol{\beta})}\right)} = O_p(1)$. For the second term, since $n \geq p \gg H^4 log(p)$, going along the lines of proof of Corollary 1.1, it can be shown that $2\sum_{j \notin O(\boldsymbol{\beta})} \overline{\Phi}\left(-\frac{n\lambda\rho'(0+)}{\|\boldsymbol{X}_j\|}\right) = o(1)$ with high probability uniformly in $\boldsymbol{\beta} \in \Theta_k^A$ with $A \gg \sqrt{\frac{log(p)}{n}}$ as long as $\lambda = \sqrt{\frac{2log(p)(1+\epsilon_p)}{(\rho'(0+))^2 n}}$ and $\epsilon_p > 0$ can be allowed to converge to 0 at slow enough rate. This completes the proof.

□

The next lemma is about properties of local minimizers of $Q(\boldsymbol{\beta}; p_\lambda)$. In particular, we are interested in the case when 0 is a local minimizer.

Lemma A.6. *Suppose p_λ satisfies condition* **(C1)**. *Then the following holds.*

1. *If $\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{n} \leq p'_\lambda(0+)$ then $\mathbf{0}$ is a local minimizer of $Q(\boldsymbol{\beta}; p_\lambda)$. Further if $\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{n} < p'_\lambda(0+)$, then 0 is unique global minimizer of $Q(\boldsymbol{\beta}; p_\lambda)$.*
2. *If $\mathbf{0}$ is a local minimizer of $Q(\boldsymbol{\beta}; p_\lambda)$, then $\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{n} \leq p'_\lambda(0+)$.*

Proof. Let $\boldsymbol{\beta} \neq \mathbf{0}$. Then,

$$\begin{aligned} Q(\boldsymbol{\beta}; p_\lambda) &= \frac{\|\boldsymbol{y}\|^2}{2n} - \frac{2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta}}{2n} + \frac{\|\boldsymbol{X}\boldsymbol{\beta}\|^2}{2n} + \sum_{j=1}^p p_\lambda(|\beta_j|) \\ &\geq \frac{\|\boldsymbol{y}\|^2}{2n} - \frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty\|\boldsymbol{\beta}\|_1}{n} + \frac{\|\boldsymbol{X}\boldsymbol{\beta}\|^2}{2n} + \sum_{j=1}^p |\beta_j| p_\lambda(\xi_j),\ \xi_j \in (0, |\beta_j|),\ j = 1, \ldots, p \\ &\geq \frac{\|\boldsymbol{y}\|^2}{2n} - \frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty\|\boldsymbol{\beta}\|_1}{n} + \frac{\|\boldsymbol{X}\boldsymbol{\beta}\|^2}{2n} + \|\boldsymbol{\beta}\|_1 p'_\lambda(0+) \end{aligned}$$

where the second to last inequality follows by Hölder's inequality and mean value theorem and the last inequality follows by condition **(C1)** since by concavity $p''_\lambda(t) \leq 0$ for $t > 0$ and hence $p'_\lambda(t)$ is a monotone decreasing function of $t > 0$. This immediately yields the proof of part (1).

The proof of part (2) of the lemma follows from KKT conditions of the optimization problem (Fan and Lv, 2011) and is omitted. □

Remark A.1. *By Lemma A.6,* $\mathbf{0}$ *is not a local minimizer if and only if* $\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{n} > p'_\lambda(0+)$. *If* $\mathbf{0}$ *is not a local minimizer, our results can be construed with respect to any choice of non-zero local minimizer of* $Q(\boldsymbol{\beta}; p_\lambda)$.

Proof of the Main Results

Proof of Proposition 1.2. It follows from Lemma A.6 that $\mathbf{0} \in \mathbb{R}^p$ is not a local minimizer of $Q_n(\boldsymbol{\beta})$ if $\|\frac{1}{\sqrt{n}}X^T y\|_\infty > \rho'(0+)\sqrt{n}\lambda$.

Also, note that by Lemma A.3, we have that

$$\mathbb{P}\left(\max_{i,j}|(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n})_{ij} - \Sigma_{ij}| > c_0\sqrt{\frac{log(p)}{n}}\right) \leq c_1 e^{-c_2 log(p)}$$

whenever $n \gg log(p)$. Hence with probability at least $1 - c_1 e^{-c_2 log(p)}$, $\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \in S_p(\gamma, 1)$ if $\gamma = O(\frac{1}{(log(p))^{2+\epsilon}})$ for some $\epsilon > 0$ provided $n \gg (log(p))^{5+2\epsilon}$ for the same $\epsilon > 0$. Also, by going along lines of proof of Corollary 1.1, it is easy to prove that the column norms of $\boldsymbol{X}$ are sharply concentrated around $\sqrt{n}$ if $n \gg H^4 log(p)$. Since, by the assumptions of the lemma, $n \gg \max((log(p))^{5+2\epsilon}, H^4 log(p))$ for any $\epsilon > 0$, we have by Lemma 11 in supplementary material of Arias-Castro et al. (2011) that $\frac{\|\frac{1}{\sqrt{n}}X^T y\|_\infty}{\sqrt{2log(p)}} \to 1$ in probability. Hence, if $\frac{\limsup \lambda}{\sqrt{\frac{2log(p)}{n}}} < \frac{1}{\rho'(0+)}$ one has that $\|\frac{1}{\sqrt{n}}X^T y\|_\infty < \rho'(0+)\sqrt{n}\lambda$ with high probability converging to 1 as $p \to \infty$. This completes the proof of the theorem. □

Proof of Theorem 1.1. (a) By the definition of NPL test as in 1.1 and the fact that $\mathbf{0} \notin C_p$ one

has by Lemma A.6,

$$\begin{aligned}
\mathbb{P}(\hat{\boldsymbol{\beta}}_\lambda \in C_p) &\leq \mathbb{P}\left(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} \geq \rho^{'}(0+)\sqrt{n}\lambda\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(|Z_j| \geq \rho^{'}(0+)\sqrt{n}\lambda) \text{ where } Z_j \sim N(0, \|\boldsymbol{X}_j\|_2^2/n) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(|Z_j| \geq \rho^{'}(0+)\sqrt{n}\lambda, B_n) + p\mathbb{P}(B_n^c) \\
&\leq p\mathbb{P}(|Z| > \rho^{'}(0+)\sqrt{n}\lambda) + p\mathbb{P}(B_n^c) \text{ where } Z \sim N(0, D_n/n) \\
&\leq pe^{-\frac{(n\lambda\rho^{'}(0+))^2}{2D_n}} + p\mathbb{P}(B_n^c) \text{ by Lemma A.1}
\end{aligned}$$

as claimed.

(b) By Lemma A.2 we have that for any $1 > \epsilon > 0$,

$$\mathbb{P}(\max_{1\leq j\leq p} \|\boldsymbol{X}_j\|_2^2 > n(1+\epsilon)) \leq 2pe^{-\frac{M\epsilon^2}{H^4}n}$$

Hence the proof follows by taking $D_n = n(1+\epsilon)$ in part (a). □

Proof of Corollary 1.1. Since the probability bound in Theorem 1.1(b) holds for any $1 > \epsilon > 0$, take $\epsilon = \delta_p/2$ for some sequence $1 > \delta_p > 0$ to be decided later. Now take $\epsilon_p = \delta_p$ in the definition of $\lambda \geq \sqrt{\frac{2(1+\epsilon_p)log(p)}{(\rho^{'}(0+))^2 n}}$. With this choice of ϵ_p, one has $\frac{n\lambda^2(\rho^{'}(0+))^2}{2(1+\delta_p/2)} \geq \frac{1+\delta_p}{1+\delta_p/2}log(p)$. Therefore, $pe^{-\frac{n\lambda^2(\rho^{'}(0+))^2}{2(1+\delta_p/2)}} \to 0$ provided $\delta_p \gg \frac{1}{log(p)}$. Also, since $n \gg H^4 log(p)$, there exists $\xi_p \to 0$ such that $n\xi_p \gg H^4 log(p)$. Therefore, taking $\epsilon_p = \delta_p = \max\{\frac{1}{\sqrt{log(p)}}, \xi_p\}$ completes the proof. □

Proof of Theorem 1.2. (a) In the following, we will not make the choice of "$\epsilon_p \to 0$ slow enough" explicit. The final rate of $\epsilon_p \to 0$ required for the validity of the proof can indeed be determined in a way similar to the proof of Corollary 1.1. Note that by Lemma A.6, $\mathbf{0}$ is not a local minimizer of $Q(\boldsymbol{\beta}; p_\lambda)$ only if $\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} > \rho^{'}(0+)\sqrt{n}\lambda$. Hence considering the test which rejects when $f(\boldsymbol{y}, \boldsymbol{X}, \hat{\boldsymbol{\beta}}_\lambda) \neq \mathbf{0}$, one has by Definition 1.1 and Corollary 1.1 that any NPL is asymptotically powerless/powerful according as $\mathbb{P}(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} > \rho^{'}(0+)\sqrt{n}\lambda) \to 1/0$. In particular, the asymptotic size is 0 by Corollary 1.1 since $n \gg H^4 log(p)$ provided $\epsilon_p \to 0$ at a slow enough rate as quantified by proof of Corollary 1.1.

For the sake of clarity, we recall that when we say all NPL tests are powerful we will mean there exists a suitable rejection region based on which the NPL test statistic is asymptotically powerful. In particular, the rejection region we demonstrate here is $(0, \infty)$. On the other hand, all NPL tests are asymptotically powerful will mean that irrespective of choice of rejection region the NPL tests are asymptotically powerless.

Now, we first show that all NPL tests are asymptotically powerless provided $A \ll \sqrt{\frac{log(p)}{n}}$. To this end note that since we are interested in the worst case power, we will demonstrate a particular $\boldsymbol{\beta}$ corresponding to H_1, along which all NPL tests will have asymptotically negligible power. In particular, fix any $O \subseteq \{1, \ldots, p\}$ with $|O| = k$ and put $\beta_j = A$ whenever $j \in O$ and $\beta_j = 0$ otherwise. According to this construction, any such $\boldsymbol{\beta} \in \Theta_k^A$. We shall show that all NPL tests have asymptotically 0 power against such $\boldsymbol{\beta}$ in the alternative.

To this end, note that by Lemma A.3, we have that

$$\mathbb{P}\left(\max_{i,j} |(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n})_{ij} - \Sigma_{ij}| > c_0\sqrt{\frac{log(p)}{n}}\right) \leq c_1 e^{-c_2 log(p)}$$

Hence with probability at least $1 - c_1 e^{-c_2 log(p)}$, $\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \in S_p(\gamma, 1)$ if $\Sigma \in S_p(\gamma, 1)$ with $\gamma \ll p^{\theta-1}$ if $n \gg H^4 p^{1-\theta}(log(p))$. Hence by our assumption, $\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \in S_p(\gamma, 1)$ with probability at least $1 - c_1 e^{-c_2 log(p)}$.

Now by Lemma 11 of supplementary material of Arias-Castro et al. (2011) we have that for $\lambda = \sqrt{\frac{2(1+\epsilon_p)log(p)}{(\rho^{'}(0+))^2 n}}$ with $\epsilon_p \to 0$ slow enough,

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} \geq \rho^{'}(0+)\sqrt{n}\lambda\right) \to 0 \tag{A.1}$$

provided $\gamma \lesssim (log(p))^{-2-\eta}$ for some $\eta > 0$. Hence by our assumptions on γ, Equation (A.1) holds. Now note that, given $\boldsymbol{X}$, for any $j = 1, \ldots, p$, $\boldsymbol{X}_j^T\boldsymbol{y} \sim N(\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}, \|\boldsymbol{X}_j\|_2^2)$. However, since $\max(H^4 log(p), p^{1-\theta} log(p))$, we have by Lemma A.3 and A.2

$$|\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}| \leq Cn(A\Delta + A(k - \Delta - 1)\gamma),\ \forall j = 1, \ldots, p$$

with probability at least $1 - e^{-c_2 log(p)} - 2pe^{-\frac{M(C-1)^2}{H^4}n}$for some constant $C > 1$.

Hence, for some constant $C_1 > 0$

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} \geq \rho^{'}(0+)\sqrt{n}\lambda\right)$$
$$\leq \mathbb{P}\left(\max_{1\leq j\leq p}|Z_j| > \sqrt{2log(p)} + \sqrt{2\epsilon_p log(p)} - C\sqrt{n}A(\Delta + (k-\Delta-1)\gamma)\right)$$
$$+ e^{-c_2 log(p)} + 2pe^{-\frac{M(C-1)^2}{H^4}n}$$

where $Z_j \sim N(0,1),\ j = 1,\ldots,p$ and $Cov(Z_j, Z_l) = \frac{\boldsymbol{X}_j^T\boldsymbol{X}_l}{n}$. Since $A \ll \sqrt{\frac{log(p)}{n}}$, $\gamma \ll p^{\theta-1}$ and $\Delta = O(1)$, we have that $\sqrt{n}A(\Delta+(k-\Delta-1)\gamma) \ll \sqrt{log(p)}$. Since $n \gg H^4 p^{1-\theta}(log(p))$, we have by equation (A.1) that $\mathbb{P}_{\boldsymbol{\beta}}(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} \geq \rho^{'}(0+)\sqrt{n}\lambda) \to 0$ if $\epsilon_p \to 0$ slow enough. Now, we show that all NPL tests are asymptotically powerful provided $A \gg \sqrt{\frac{log(p)}{n}}$. In particular, we show that for any configuration of the alternative $\boldsymbol{\beta}$ since the minimum signal strength is at least A, we will have $\mathbb{P}_{\boldsymbol{\beta}}(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} \geq \rho^{'}(0+)\sqrt{n}\lambda) \to 1$ uniformly in any configuration of the alternative. Take any particular $\boldsymbol{\beta} \in \Theta_k^A$ and let $O(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta} \neq 0\}$ with $|O(\boldsymbol{\beta})| = k$ and $\min_{j\in O(\boldsymbol{\beta})}|\beta_j| \geq A$. Similar to previous argument, given $\boldsymbol{X}$, we have for any $j = 1,\ldots,p$, $\boldsymbol{X}_j^T\boldsymbol{y} \sim N(\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}, \|\boldsymbol{X}_j\|_2^2)$. Once again, we have by Lemma A.3 and A.2

$$|\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}| \geq Cn(A\Delta - A(k-\Delta-1)\gamma),\ \forall j = 1,\ldots,p$$

with probability at least $1 - e^{-c_2 log(p)} - 2pe^{-\frac{M(C-1)^2}{H^4}n}$for some constant $C > 1$. Note that this independent of the configuration of the alternative. Hence, for some constant $C_1 > 0$

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} > \rho^{'}(0+)\sqrt{n}\lambda\right)$$
$$\geq \mathbb{P}\left(\max_{1\leq j\leq p}|Z_j| \leq -\sqrt{2log(p)} - \sqrt{2\epsilon_p log(p)} + C\sqrt{n}A(\Delta - (k-\Delta-1)\gamma)\right)$$
$$- e^{-c_2 log(p)} - 2pe^{-\frac{M(C-1)^2}{H^4}n}$$

where Z_j's are as defined earlier. However, $A \gg \sqrt{\frac{log(p)}{n}}$ along with conditions on γ, Δ, n together with equation (A.1) implies the result if $\epsilon_p \to 0$ slow enough. This concludes the proof of part (i).

The proof of part (ii) follows directly using the argument of part (a) and proof of Theorem 5 of Arias-Castro et al. (2011) since $\epsilon_p < (1-\sqrt{1-\theta})^2$ for large enough p if $\epsilon_p \to 0$ since

$\gamma \ll p^{\theta-1}$ implies that $\gamma^2 \ll p^{\theta-1}(log(p))^{-3}$ as required by Theorem 5 of Arias-Castro et al. (2011).

(b) The proof follows along similar lines as the proof of detection limits of ANOVA in Arias-Castro et al. (2011) for Proposition 3. By proof similar to that of Corollary 1.1 and Theorem 1.2(a), the assumptions on $\boldsymbol{X}$, as needed for the validity of the results in Arias-Castro et al. (2011), hold with high probability under the assumptions on n, p, H and Σ since $n \gg \max(H^4 log(p), (log(p))^3)$. We omit the details here. □

Proof of Theorem 1.3. First let the $0 < \theta \leq \frac{1}{2}$. Under the assumptions of the theorem, the proof follows along lines of proof of Proposition 3 in Arias-Castro et al. (2011) and $t_p(0)$ can be taken as $t_p(0) = p + \eta_p\sqrt{p}$ for a slow enough divergent sequence $\eta_p > 0$. By proof similar to that of Corollary 1.1 and Theorem 1.2, the assumptions on $\boldsymbol{X}$, as needed for the validity of the results in Arias-Castro et al. (2011), hold with high probability under the assumptions on n, p, H and Σ since $n \gg \max(H^4 log(p), (log(p))^3)$. We omit the details here. Now let us prove that for $\theta > \frac{1}{2}$. We will show that by rejecting when $\|\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\|_2^2 > t_p$ where $t_p >$ is slowly diverging and $\lambda = \sqrt{\frac{2(1+\epsilon)log(p)}{(\rho'(0+))^2 n}}$, the test is asymptotically powerful. The type I error converges to 0 of this test is asymptotically negligible by Theorem 1.1. So we only need to show that the type II error converges to 0 uniformly in $\boldsymbol{\beta} \in \Theta_k^A$. The crucial ingredient of the proof is noting that under the assumptions of the theorem, since $\rho^{'}(0+) = 1$ for Lasso penalty, one has by Bühlmann and Van De Geer (2011) that for any fixed $\epsilon > 3$

$$\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})\|_2^2/n = O_p(\frac{klog(p)}{n}) \tag{A.2}$$

uniformly in $\boldsymbol{\beta} \in \Theta_k^A$ provided $k \ll \frac{n}{log(p)}$ which holds by our assumption on n and p. Hence,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\beta}}\left(\|\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\|_2^2 > t_p\right) \geq \mathbb{P}_{\boldsymbol{\beta}}\left(\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})\|_2^2 \leq \|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - t_p\right) \\
&\geq \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})\|_2^2}{n} \leq s_{min}(\frac{\boldsymbol{X}_O^T\boldsymbol{X}_O}{n})\|\boldsymbol{\beta}\|_2^2 - \frac{t_p}{n}\right) \text{ where } O = \{j : \beta_j \neq 0\} \\
&\geq \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta})\|_2^2}{n} \leq \xi_p s_{min}(\frac{\boldsymbol{X}_O^T\boldsymbol{X}_O}{n})\frac{klog(p)}{n} - \frac{t_p}{n}\right) \text{ where } \xi_p \to \infty
\end{aligned} \tag{A.3}$$

where the last line follows since $\boldsymbol{\beta} \in \Theta_k^A$. Now arguing as proof of Lemma A.4 as Ver-

shynin (2010), $s_{min}\left(\frac{\boldsymbol{X}_O^T\boldsymbol{X}_O}{n}\right) = O\left(s_{min}(\Sigma_O) + \sqrt{\frac{k}{n}}\right)$ with probability at least $1 - 2e^{-\frac{k}{H^4}}$. Hence, by assumptions on Σ, H and equations (A.2) and (A.3) we have that $\mathbb{P}_{\boldsymbol{\beta}}(\|\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\|_2^2 > t_p) \to 1$ uniformly in $\boldsymbol{\beta} \in \Theta_k^A$ whenever t_p is such that $t_p \ll p^\delta log(p)$ for all $\delta > 0$. This completes the proof of the theorem. □

Proof of Theorem 1.4. First note that the asymptotic sizes of the tests are 0 by Corollary 1.1 since $n \gg H^4 log(p)$ provided $\epsilon_p \to 0$ at a slow enough rate as quantified by proof of Corollary 1.1. The proof then follows directly from Lemma A.5(b).

□

Proof of Proposition 1.3. By simple calculations since $\mathrm{KL}(\mathbb{P}_0|\mathbb{P}_{\boldsymbol{\beta}}) = \frac{n}{2}\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta} \leq \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 s_{\max}(\Sigma_{O(\boldsymbol{\beta})})$ and the proof follows. □

Proof of Theorem 1.5. By Lemma A.6, we have that

$$\mathbb{P}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_\lambda \in C_p) \leq \mathbb{P}_{\boldsymbol{\beta}}(\frac{\|\boldsymbol{X}^T\boldsymbol{y}\|_\infty}{\sqrt{n}} > \sqrt{n}\lambda\rho^{'}(0+))$$

Now, given $\boldsymbol{X}$, $\boldsymbol{X}_j^T\boldsymbol{y} \sim N(\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}, \|\boldsymbol{X}_j\|_2^2)$. Letting Let $O = \{j : \beta_j \neq 0\}$, we have by Cauchy Schwarz Inequality the following hold uniformly in all $\boldsymbol{\beta} \in R_k^p$

$$\begin{aligned}\frac{|\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}|}{\sqrt{n}} &\leq \|\boldsymbol{X}_j\|\|\boldsymbol{\beta}\|\sqrt{s_{\max}(\frac{\boldsymbol{X}_O^T\boldsymbol{X}_O}{n})} \\ &\leq 2\sqrt{n}\|\boldsymbol{\beta}\|\sqrt{s_{\max}(\Sigma_O) + \sqrt{\frac{k}{n}}C_H} \text{ w.p. } \geq 1 - 2ke^{-\frac{16M^2}{H^4}n} - 2e^{-c_H k}\end{aligned}$$

where the last inequality follows by Lemmas A.4 and A.2 with $M > 0$ a constant and $C_H, c_H > 0$ only depend on H. It can be argued along the lines of proof of Lemma A.4 as in Vershynin (2010) that C_H can be taken to be $O(1)$ and $c_H = O(1/H^4)$ for the sake of our purposes. Since , $n\|\boldsymbol{\beta}\|^2 = O(1)$ there exists absolute constants $c, C > 0$ such that with probability at least $1 - 2ke^{-\frac{16M^2}{H^4}n} - 2e^{-ck/H^4}$ one has

$$\frac{|\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}|}{\sqrt{n}} \leq C\nu_p, \ \forall j$$

where $\nu = O(\sqrt{s_{\max}(\Sigma_O) + \sqrt{\frac{k}{n}}})$. Hence, by similar arguments as in proof of Theorem 1.2

and similarly defined Z_j's, we have

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_\lambda \in C_p) &\leq P(\max_{1\leq j\leq p}|Z_j| > \sqrt{2(1+\epsilon_p)log(p)} - C\nu_p) \\ &\leq pe^{-\frac{(\sqrt{2(1+\epsilon_p)log(p)}-C\nu_p)^2}{2}} + 2ke^{-\frac{16M^2}{H^4}n} + 2e^{-ck/H^4} = O(p^{-\eta_p})\end{aligned}$$

for some $\eta_p \to 0$ provided $\epsilon_p \to 0$ slow enough. The second inequality follows by Lemma A.1 and the last line holds since $\max(s_{\max}(\Sigma_O), H) \ll log(p)$ and therefore along with the fact that $\frac{k}{n} \ll log^2(p)$ one also has $n \gg H^4(log(p))^{1+\delta}$ for all fixed $\delta > 0$.

Now, we note that, since the power of the LR test is an increasing function of the non-centrality parameter of the chi-square test, we have by Berry-Esseen Theorem and Lemma A.1 that for some constant $C_1, C_2 > 0$,

$$\begin{aligned}\mathbb{P}(T_n^{LR} > t_p) &\geq C_1\phi(\frac{t_p - r}{\sqrt{r}})\frac{\sqrt{r}}{t_p - r} \\ &\geq C_2\frac{e^{-\delta_p^2}}{\delta_p}\end{aligned}$$

where $r = p \wedge n$ and $\delta_p = \frac{t_p - r}{\sqrt{r}}$. Now, since $log(p) \ll n$, we have by the assumptions and part (a) of the theorem, that there exists a $\eta > 0$ such that $\mathbb{P}(\hat{\boldsymbol{\beta}}_\lambda \in C_p) \leq p^{-\eta_p}$. Now, choosing $\epsilon_p \to 0$ slow enough, the rate of convergence of η_p to 0 can be made arbitrarily slow. Hence for $t_p \to \infty$ chosen so that $\frac{e^{-\delta_p^2}}{\delta} \gg p^{-\gamma}$ for all $\gamma > 0$, we have the required result. □

Appendix B

Proofs for Chapter 2

Notations

We begin by briefly summarizing notation. We recall the definition of our chosen prior π for the sake of completeness. We choose π to be uniform over all k sparse subsets of $\mathbb{R}^p$ with signal strength either A or $-A$. Let $M(k,p)$ be the collection of all subsets of $\{1,\ldots,p\}$ of size k. For each $m \in M(k,p)$, let $\xi^m = (\xi_j)_{j\in m}$ be a sequence of independent Rademacher random variables taking values in $\{+1,-1\}$ with equal probability. Given $A > 0$ for testing (2.4), a realization from the prior distribution π on $\mathbb{R}^p$ can be expressed as $\beta_{\xi,m} = \sum_{j\in m} A\xi_j e_j$, where $(e_j)_{j=1}^p$ is the canonical basis of $\mathbb{R}^p$ and m is uniformly chosen from $M(k,p)$. In the following we will define m_1, m_2 to be two independent draws at random from $M(k,p)$ and $\xi_1 = (\xi_1^j)_{j\in m}, \xi_2 = (\xi_2^j)_{j\in m}$ the corresponding draws of a sequence of Radamacher random variables. Further we denote by m_3 and m_4 the set valued random variables $m_3 := \{j \in m_1 \cap m_2 : \xi_1^j = \xi_2^j\}$ and $m_4 := \{j \in m_1 \cap m_2 : \xi_1^j = -\xi_2^j\}$. Also ϕ, Φ and $\overline{\Phi}$ denote the standard normal pdf, cdf and survival functions respectively. We let $\mathrm{Hypergeometric}(N,m,n)$ denote the hypergeometric distribution counting the number of red balls in n draws from an urn containing m red balls out of N. Also throughout C will denote generic positive constants whenever necessary.

Preliminary Lemmas

We will use the following results many times and hence present them as useful lemmas. The first result compares the hypergeometric distribution with a related binomial distri-

bution, which is in general simpler to work with.

Lemma B.1. *If* $W \sim \text{Hypergeometric}(N, m, n)$ *and* $Y \sim \text{Bin}(n, \frac{m}{N-m})$ *then* W *is stochastically smaller than* Y*, i.e.,* $\mathbb{P}(Y \geq t) \geq \mathbb{P}(X \geq t)$ *for all* $t \in \mathbb{R}$*. Moreover this implies that for any non-decreasing function* g *one has* $\mathbb{E}(g(W)) \leq \mathbb{E}(g(Y))$*.*

Proof. The proof can be found in Arias-Castro et al. (2011) and follows by noting that if the balls are picked one by one without replacement, then at each stage, the probability of selecting a red ball is smaller than $m/(N-m)$. □

The next result presents an inequality about the tail probability of a binomial random variable (Carter and Pollard, 2004)

Lemma B.2. *Let* $X \sim \text{Bin}(n, \frac{1}{2})$ *with* $n \geq 28$*. Define*

$$\gamma(\epsilon) = \frac{(1+\epsilon)log(1+\epsilon) + (1-\epsilon)log(1-\epsilon) - \epsilon^2}{2\epsilon^4} = \sum_{l=0}^{\infty} \frac{\epsilon^2 l}{(2l+3)(2l+4)},$$

an increasing function. Suppose $\frac{n}{2} < k' \leq n-1$*. Define* $\epsilon = (2K-N)/N$*, where* $K = k'-1$ *and* $N = n-1$*. Then there exists a* λ_n *such that* $\frac{1}{12n+1} < \lambda_n < \frac{1}{12n}$ *and a constant* C *such that*

$$\mathbb{P}(X \geq k') = \overline{\Phi}(\epsilon\sqrt{N})e^{A_n(\epsilon)}$$

where

$$A_n(\epsilon) = -N\epsilon^4\gamma(\epsilon) - \frac{1}{2}log(1-\epsilon^2) - \lambda_{n-k} + r_{k'}$$

and

$$-ClogN \leq Nr_{k'} \leq C$$

for all ϵ *corresponding to the range* $\frac{n}{2} < k' \leq n-1$*.*

The next lemma shows that any random draw of a subset of size k from $\{1, \ldots, p\}$ can have at most one element in each block. The proof of the lemma is similar to the proof of Lemma A.8 of Hall and Jin (2010) and is omitted.

Lemma B.3. *Let $t_1 < t_2 < \ldots < t_k$ be k distinct indices randomly sampled from $\{1, \ldots, p\}$ without replacement. Then for any $1 \leq Q \leq p$ we have $\mathbb{P}(\min_{1 \leq i \leq k-1} |t_{i+1} - t_i| \leq Q) \leq Qk(k+1)/p$.*

The next Lemma is tailored towards controlling the contribution of the i^{th} row in the expression for $\mathbb{E}_0(L_\pi^2)$.

Lemma B.4. *Suppose for the i^{th} row of $\boldsymbol{X}$ one has $|S_i| \leq Q$ and that the elements of $\boldsymbol{X}$ are bounded by M in absolute value. Then for any $\beta, \beta' \sim \pi$,*

$$\theta(\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(\boldsymbol{x}_i^t\boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(-\boldsymbol{x}_i^t\boldsymbol{\beta}') \leq \theta^2(QMA) + \theta^2(-QMA).$$

where θ is the distribution function of a symmetric random variable, i.e., θ satisfies Equation 2.3.

Proof. We begin by noting that for any i,

$$\theta(\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(\boldsymbol{x}_i^t\boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(-\boldsymbol{x}_i^t\boldsymbol{\beta}') \leq \sup_{s_1, s_2 \in [-MQ, MQ]} \theta(s_1 A)\theta(s_2 A) + \theta(-s_1 A)\theta(-s_2 A).$$

Hence by symmetry of the above supremum in s_1, s_2 and using the fact that $\theta(z) + \theta(-z) = 1$ for all w, we have that

$$\theta(\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(\boldsymbol{x}_i^t\boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(-\boldsymbol{x}_i^t\boldsymbol{\beta}') \leq \max_{s \in [0, MQ]} (\theta(sA))^2 + (1 - \theta(sA))^2.$$

Now noting that $(1-w)^2 + w^2$ is an increasing function of w for $w \geq \frac{1}{2}$ and using the fact that $\theta(sA) \geq \frac{1}{2}$ for $s \geq 0$, we have the desired result. □

Proof of Main Results

Proof of Theorem 2.1 . We will produce one prior $\pi_0 \sim \pi$ for which the theorem holds. Hence, for any other $\pi^* \sim \pi$, since one also has $\pi^* \sim \pi_0$ we have the result holding by a similar proof. We begin by noting that

$$\theta(\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(\boldsymbol{x}_i^t\boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(-\boldsymbol{x}_i^t\boldsymbol{\beta}') \leq 1 \text{ for all } i, \boldsymbol{\beta}, \boldsymbol{\beta}' \tag{B.1}$$

The proof of (B.1) follows from noting that for any two real numbers w_1, w_2, one has by symmetry $\theta(w_1)\theta(w_2) + \theta(-w_1)\theta(-w_2) \leq \sup_{w \in \mathbb{R}} [2\theta^2(w) - 2\theta(w) + 1]$. Since θ is a distribution

function of a symmetric random variable as posed by equation (2.2), it is easy to show that $2\theta^2(w) - 2\theta(w) + 1$ is an increasing function of w. Hence we have that the supremum equals 1 and thus proving (B.1). Now, recall that it suffices to bound from below the second moment $\mathbb{E}_0(L_\pi^2)$ where by Fubini's Theorem

$$\begin{aligned}
\mathbb{E}_0(L_\pi^2) &= 2^n \iint \prod_{i=1}^{n} \Big[\theta(\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(\boldsymbol{x}_i^t\boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(-\boldsymbol{x}_i^t\boldsymbol{\beta}')\Big] d\pi(\boldsymbol{\beta})d\pi(\boldsymbol{\beta}') \\
&\leq \iint 2^{n-\sum_{i=1}^{n} \mathbf{I}(\min\{|m_1\cap S_i|,|m_2\cap S_i|\}=0)} d\pi(\boldsymbol{\beta})d\pi(\boldsymbol{\beta}') \\
&= \iint 2^{\sum_{i=1}^{n} \mathbf{I}(\min\{|m_1\cap S_i|,|m_2\cap S_i|\}>0)} d\pi(\boldsymbol{\beta})d\pi(\boldsymbol{\beta}'). \qquad \text{(B.2)}
\end{aligned}$$

The inequality in the second to last line above follows from noting that, when i is such that one of $S_i \cap m_1$ or $S_i \cap m_2$ is empty, then the integrand $\theta(\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(\boldsymbol{x}_i^t\boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t\boldsymbol{\beta})\theta(-\boldsymbol{x}_i^t\boldsymbol{\beta}') = \frac{1}{2}$, whereas for any other i, the integrand is less than or equal to 1 by (B.1). Applying Lemma B.3 we obtain that when $\alpha > \frac{1}{2}$, *i.e.*, $k = p^{1-\alpha} \ll \sqrt{p}$, it makes negligible difference by restricting π to $R_p = \{\{t_1, \ldots, t_k\}, \min_{1\leq i\leq k-1} |t_{i+1} - t_i| > \sigma_p\}$ where by assumption σ_p is such that $\sigma_p \ll p^\epsilon$ for all $\epsilon > 0$. If we denote this restricted prior by π_0, then we have $\pi_0 \sim \pi$ and $R_{\pi_0} = R_p$. Now by elementary combinatorics,

$$|R_{m_1}^N(\sigma_p)| \lesssim \binom{k}{N}(2\sigma_p)^N \binom{p-N}{k-N} \leq \binom{k}{N}(2\sigma_p)^N \binom{p}{k-N}.$$

Also by direct calculation,

$$\frac{\binom{k}{N}\binom{p}{k-N}}{\binom{p}{k}} = \frac{1}{N!}\left(\frac{k!}{(k-N)!}\right)^2 \frac{(p-k)!}{(p-k+N)!} \lesssim \frac{1}{N!}\left(\frac{k^2}{p}\right)^N.$$

Hence from (B.2) and assumption of the Theorem we have that

$$\begin{aligned}
\mathbb{E}_0(L_\pi^2) &\leq \binom{p}{k}^2 \sum_{m_1 \in R_{\pi_0}} \sum_{N=0}^{k} \sum_{m_2 \in R_{m_1}^N(\sigma_p)} 2^{\sum_{i=1}^{n} \mathbf{I}(\min\{|m_1 \cap S_i|, |m_2 \cap S_i|\} > 0)} (1 + o(1)) \\
&\leq \binom{p}{k}^2 \sum_{m_1 \in R_{\pi_0}} \sum_{N=0}^{k} \sum_{m_2 \in R_{m_1}^N(\sigma_p)} 2^{N\delta_p} (1 + o(1)) \\
&\lesssim \binom{p}{k} \sum_{m_1 \in R_{\pi_0}} \sum_{N=0}^{\infty} \frac{2\frac{k^2}{p}\sigma_p {2^{\delta_p}}^N}{N!} (1 + o(1)) \\
&= \binom{p}{k} \sum_{m_1 \in R_{\pi_0}} e^{2\frac{k^2}{p}\sigma_p 2^{\delta_p}} (1 + o(1)) \\
&= e^{2\frac{k^2}{p}\sigma_p 2^{\delta_p}} (1 + o(1))
\end{aligned}$$

Since σ_p is a poly-logarithmic factor of p and $k = p^{1-\alpha}$ with $\alpha > \frac{1}{2}$, we have that $\delta_p \ll log(p)$ implies that $\mathbb{E}_0(L_\pi^2) = 1 + o(1)$. Hence all tests are asymptotically powerless as required. □

Proof of Theorem 2.2. The proof relies on verifying the assumptions and conditions of Theorem 2.1. To begin with we produce a prior that is equivalent to π as follows. Let π_0 be the restriction of π to
$R_p = \{\{t_1, \ldots, t_k\}, \min_{1 \leq i \leq k-1} |t_{i+1} - t_i| > \sigma_p\}$ and let $\pi_{0,1}$ be the restriction of π_0 to $(\bigcup_{i \notin \Omega} S_i)^c$ where $\sigma_p \geq 2l^*$ is such that $\sigma_p \ll p^\epsilon$ for all $\epsilon > 0$. We note that such a σ_p can be found since we have by assumption $l^* \ll p^\epsilon$ for all $\epsilon > 0$. Since $k = p^{1-\alpha}$ with $\alpha > \frac{1}{2}$, by Lemma B.3 and the fact $|\bigcup_{i \notin \Omega} S_i| \ll p$ we have that $\pi_{0,1} \sim \pi_0 \sim \pi$. Since any draw from $\pi_{0,1}$ does not intersect with S_i with $i \notin \Omega$, we have that

$$\sum_{i=1}^{n} \mathbf{I}(\min\{|m_1 \cap S_i|, |m_2 \cap S_i|\} > 0) = \sum_{i \in \Omega} \mathbf{I}(\min\{|m_1 \cap S_i|, |m_2 \cap S_i|\} > 0).$$

Let m_1 and m_2 be two independent draws from $\pi_{0,1}$ with $m_2 \in \tilde{R}_p^N(2\sigma_p)$. We have that there must exist exactly N blocks $T_{j_1}, \ldots, T_{j_N}$ which have elements from m_1 and m_2 σ_p-mututallyclose. In the rest of the $M - N$ blocks there is either no element of m_1 or no element of m_2. Hence the total number of rows corresponding to $\mathbf{I}(\min\{|m_1 \cap S_i|, |m_2 \cap$

$S_i|\} > 0)$ equals $\sum_{l=1}^{N} c_{j_l} \leq Nc^*$. Hence we have

$$\sum_{i=1}^{n} \mathbf{I}(\min\{|m_1 \cap S_i|, |m_2 \cap S_i|\} > 0) \leq Nc^*$$

for the prior $\pi_{0,1} \sim \pi$ and all m_1, m_2 drawn from $\pi_{0,1}$ with $m_2 \in \tilde{R}_p^N(2\sigma_p)$. So by Theorem 2.1, we have that if $c^* \ll log(p)$ then all tests are asymptotically powerless. □

Proof of Theorem 2.3 . The proof follows by arguments similar to that of Theorem 2.2 and hence is omitted. □

Proof of Theorem 2.4. Since for each $t > 0$, $W_p(t)$ is a normalized mean of i.i.d random variables, by the union bound and Chebyshev's Inequality,

$$\begin{aligned}\mathbb{P}(\mathsf{T}_{\mathrm{HC}} > log(p)) &\leq \sum_{t \in [1, \sqrt{3log(p)}] \cap \mathbb{N}} \mathbb{P}(W_p(t) > log(p)) \\ &\leq 2\sqrt{3log(p)} \frac{1}{(log(p))^2} = o(1)\end{aligned}$$

□

Proof of Theorem 2.5. The proof of this theorem follows techniques similar to the proof of Theorem 2.6. However, this can be proved from much simpler combinatorial arguments and hence we provide the proof for the sake of interest. We divide the proof of the theorem into three paragraphs, namely, two-sided alternatives, one-sided alternative for sparse regime and one-sided alternative for dense regime, which correspond to the three parts of the theorem.

Proof of Part(1): Two-Sided Alternatives

We do the proof for logistic regression for the sake of clarity and note that the proof for general binary regression is exactly same, because the proof only uses the fact $\theta(x) + \theta(-x) = 1$ for the logistic distribution function which is symmetric. Using Remark 2.1, the proof also holds for problem 2.13. Although the following proof is carried out in the usual way of analyzing the second moment of the likelihood ratio as in the proof of

Theorem 2.7, here we provide a more direct combinatorial proof.

For logistic regression, we have

$$L_\pi = 2^p \int \prod_{j=1}^{p} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} d\pi(\boldsymbol{\beta}) = 2^p . \frac{1}{2^k} \frac{1}{\binom{p}{k}} \sum_{m,\xi} \prod_{j=1}^{p} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}}.$$

Take any instance of (m, ξ), say, $m = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, p\}$ and $\xi = \{\sigma_1, \ldots, \sigma_k\}$, $\sigma_l \in \{-1, 1\}$, $l = 1, \ldots, k$. Then

$$\prod_{j=1}^{p} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} = \left(\frac{1}{2}\right)^{p-k} \prod_{j \in m} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} \, .$$

Hence,

$$\begin{aligned} L_\pi &= \frac{1}{\binom{p}{k}} \sum_{m,\xi} \prod_{j \in m} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} \\ &= \frac{1}{\binom{p}{k}} \sum_{\{i_1,\ldots,i_k\} \subseteq \{1,\ldots,p\}} \sum_{r=0}^{k} \sum_{\{j_1,\ldots,j_r\} \subseteq \{i_1,\ldots,i_k\}} \frac{e^{Ay_{j_1}} \cdots e^{Ay_{j_r}} e^{A(1-y_{j_{r+1}})} \cdots e^{A(1-y_{j_k})}}{(1+e^A)^k} \end{aligned}$$

where $\{j_{r+1}, \ldots, j_k\} = \{i_1, \ldots, i_k\} \cap \{j_1, \ldots, j_r\}^c$. Now we claim that for any subset $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, p\}$,

$$\sum_{r=0}^{k} \sum_{\{j_1,\ldots,j_r\} \subseteq \{i_1,\ldots,i_k\}} \frac{e^{Ay_{j_1}} \cdots e^{Ay_{j_r}} e^{A(1-y_{j_{r+1}})} \cdots e^{A(1-y_{j_k})}}{(1+e^A)^k} = 1$$

for any sample $(y_1, \ldots, y_p)$. To see this, given a sample $(y_1, \ldots, y_p)$ and a subset $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, p\}$, the number of times the summand equals $\frac{e^{Al}}{(1+e^A)^k}$ is $\binom{k}{l}$ for any $l = 0, 1, \ldots, k$ (because any y_j is either 0 or 1)and this exhausts the sum. Hence the total equals

$$\sum_{r=0}^{k} \sum_{\{j_1,\ldots,j_r\} \subseteq \{i_1,\ldots,i_k\}} \frac{e^{Ay_{j_1}} \ldots e^{Ay_{j_r}} e^{A(1-y_{j_{r+1}})} \ldots e^{A(1-y_{j_k})}}{(1+e^A)^k} = \sum_{l=0}^{k} \frac{\binom{k}{l} e^{Al}}{(1+e^A)^k} = 1$$

as claimed. Hence $L_\pi = 1$ for any sample. Hence by noting that for any test T, $\mathrm{Risk}_\pi(T) \geq 1 - \frac{1}{2}\mathbb{E}_0|L_\pi - 1| \geq 1$ we have that all tests are powerless.

Proof of Part(2a): One-Sided Alternatives, Dense Regime

We divide our proof into that of lower bound and upper bound.

Proof of Lower Bound We will do the proof for general binary regression i.e. $\mathbb{E}(Y_j) = \theta(\beta_j)$, $j = 1, \ldots, p$ where θ is any distribution function of a symmetric random variable, *i.e.*, $\theta(x) + \theta(-x) = 1$ for all x and $\theta \in BC^1(0)$. Hence, by Remark 2.1, the proof for lower bound in problem 2.13 follows. Note that one can express $\mathbb{E}_0(L_\pi^2)$ as follows:

$$\begin{aligned}
\mathbb{E}_0(L_\pi^2) &= \mathbb{E}_{m_1,m_2,\xi_1,\xi_2}[\{4\theta^2(A) - 4\theta(A) + 2\}^{|m_1 \cap m_2| \frac{|\xi_1+\xi_2|}{2}} \{4\theta(A)\theta(-A)\}^{|m_1 \cap m_2| \frac{|\xi_1-\xi_2|}{2}}] \\
&= \frac{1}{2}\mathbb{E}_{m_1,m_2}[\{4\theta^2(A) - 4\theta(A) + 2\}^{|m_1 \cap m_2|} + \{4\theta(A)\theta(-A)\}^{|m_1 \cap m_2|}] \\
&\leq \mathbb{E}_{m_1,m_2}[\{4\theta^2(A) - 4\theta(A) + 2\}^{|m_1 \cap m_2|}].
\end{aligned}$$

The last line is true because $4\theta^2(A) - 4\theta(A) + 2 \geq \max\{1, 4\theta(A)\theta(-A)\}$. Now we note that $|m_1 \cap m_2| \sim \text{Hypergeometric}(p, k, k)$ which is stochastically smaller than $\text{Bin}(k, \frac{k}{p-k})$ by Lemma B.1. Since $4\theta^2(A) - 4\theta(A) + 2 \geq \max\{1, 4\theta(A)\theta(-A)\}$ one has that for $Z \sim \text{Bin}(k, \frac{k}{p-k})$,

$$\begin{aligned}
\mathbb{E}_0(L_\pi^2) &\leq \mathbb{E}_{m_1,m_2}[\{4\theta^2(A) - 4\theta(A) + 2\}^{|m_1 \cap m_2|}] \leq \mathbb{E}_Z[\{4\theta^2(A) - 4\theta(A) + 2\}^Z] \\
&= \left[\frac{p-2k}{p-k} + \frac{k}{p-k}(4\theta^2(A) - 4\theta(A) + 2)\right]^k = \left[1 + \frac{\frac{k^2}{p-k}(2\theta(A) - 1)^2}{k}\right]^k \\
&= \left[1 + \frac{\frac{k^2}{p-k}(2A\theta'(0) + O(A^2))^2}{k}\right]^k = 1 + o(1)
\end{aligned}$$

since $p^{1-2\alpha}A \to 0$

Proof of Upper Bound The proof is similar to the proof of upper bound in Theorem 2.6 in the main text and is based on comparing second moment and variance of the test statistic under the alternative. Hence we skip the details of the proof. However we provide another proof here by showing that the test based on $\sum y_i$ is actually the most powerful Bayes test by showing that the likelihood ratio is a function of $\sum y_i$. The argument goes

as follows for the logistic regression. The proof for general binary regression follows by similar arguments due to symmetry of the link function.

To this end, note that

$$\begin{aligned} L_\pi &= 2^p \int \prod_{j=1}^{p} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} d\pi(\boldsymbol{\beta}) \\ &= 2^p \frac{1}{2} \frac{1}{\binom{p}{k}} \sum_{m,\xi} \prod_{j=1}^{p} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} \end{aligned}$$

Take any instance of (m, ξ), say, $m = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, p\}$ and $\xi = 1$.Then

$$\prod_{j=1}^{p} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} = \left(\frac{1}{2}\right)^{p-k} \prod_{l=1}^{k} \frac{e^{A y_{j_l}}}{1+e^{A}} = \left(\frac{1}{2}\right)^{p-k} \frac{e^{A\sum_{l=1}^{k} y_{j_l}}}{(1+e^{A})^k}$$

For $\xi = -1$ one has

$$\prod_{j=1}^{p} \frac{e^{\beta_j y_j}}{1+e^{\beta_j}} = \left(\frac{1}{2}\right)^{p-k} \prod_{l=1}^{k} \frac{e^{-A y_{j_l}}}{1+e^{-A}} = \left(\frac{1}{2}\right)^{p-k} \frac{e^{A\sum_{l=1}^{k} (1-y_{j_l})}}{(1+e^{A})^k}$$

Hence,

$$L_\pi = \frac{2^k}{2\binom{p}{k}} \sum_{\{j_1,\ldots,j_k\} \subseteq \{1,\ldots,p\}} \frac{e^{A\sum_{l=1}^{k} y_{j_l}} + e^{A\sum_{l=1}^{k} (1-y_{j_l})}}{(1+e^{A})^k} \tag{B.3}$$

Now suppose there are two samples $\boldsymbol{y}^{(1)}$ and $\underline{\boldsymbol{y}^{(2)}}$ both have which have exactly r 1's in positions say $\{i_1^{(1)}, \ldots, i_r^{(1)}\}$ and $\{i_1^{(2)}, \ldots, i_r^{(2)}\}$ respectively. Now each $\{j_1, \ldots, j_k\} \subseteq \{1, \ldots, p\}$ in the summand in (B.3) can be partitioned as below:

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(1)}, \ldots, i_r^{(1)}\}$ in 0 position: $n_0^{(1)}$ ways

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(1)}, \ldots, i_r^{(1)}\}$ in 1 position: $n_1^{(1)}$ ways

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(1)}, \ldots, i_r^{(1)}\}$ in 2 positions: $n_2^{(1)}$ ways

$\vdots$

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(1)}, \ldots, i_r^{(1)}\}$ in r positions: $n_r^{(1)}$ ways

or

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(2)}, \ldots, i_r^{(2)}\}$ in 0 position: $n_0^{(2)}$ ways

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(2)}, \ldots, i_r^{(2)}\}$ in 1 position: $n_1^{(2)}$ ways

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(2)}, \ldots, i_r^{(2)}\}$ in 2 positions: $n_2^{(2)}$ ways

$\vdots$

$\{j_1, \ldots, j_k\}$ intersect $\{i_1^{(2)}, \ldots, i_r^{(2)}\}$ in r positions: $n_r^{(2)}$ ways (say)

Now it is easy to see that $n_l^{(1)} = n_l^{(2)}$ for $l = 0, 1, \ldots, r$ and for each of these partitions the summand is the same value. Hence the total sum is same provided two same sample has the same number of 1's and therefore proving the claim.

Proof of Part(2b): One-Sided Alternatives, Sparse Regime

We give the proof for logistic regression and note that the proof for general binary regression is exactly same because the proof uses only the fact $\theta(x) + \theta(-x) = 1$ for the logistic distribution function which is symmetric. Using Remark 2.1, the proof also holds for problem 2.13. Although the following proof can be proved in the usual way of analyzing the second moment of the likelihood ratio, here we provide a more combinatorial proof without using Lemma B.1.

Note that we have by Fubini's Theorem,

$$\mathbb{E}_0(L_\pi^2) = 2^p . \frac{1}{4} . \frac{1}{\binom{p}{k}^2} \sum_{(m_1,\xi_1),(m_2,\xi_2)} \left(\frac{1}{2}\right)^{|m_1 \Delta m_2|} \left\{\frac{1+e^{2A}}{(1+e^A)^2}\right\}^{|m_1 \cap m_2| \frac{\xi_1+\xi_2}{2}} \left\{\frac{2e^A}{(1+e^A)^2}\right\}^{|m_1 \cap m_2| \frac{|\xi_1-\xi_2|}{2}}$$

$$= 2^p . \frac{1}{4} . \frac{1}{\binom{p}{k}^2} \sum_{r=0}^{k} \sum_{(m_1,\xi_1),(m_2,\xi_2):|m_1 \cap m_2|=r} \left(\frac{1}{2}\right)^{p-r} \left\{\frac{1+e^{2A}}{(1+e^A)^2}\right\}^{r \frac{\xi_1+\xi_2}{2}} \left\{\frac{2e^A}{(1+e^A)^2}\right\}^{r \frac{|\xi_1-\xi_2|}{2}}$$

where $(m_1, \xi_1), (m_2, \xi_2)$ are i.i.d.

First consider $r = 0$. Then $m_1 \cap m_2 = \Phi$. The number of such tuples (m_1, m_2) is $\binom{p}{k}\binom{p-k}{k}$. For each such $\binom{p}{k}\binom{p-k}{k}$ combinations of $(m_1, \xi_1), (m_2, \xi_2)$ the summand above equals $(\frac{1}{2})^p$. Hence total $= \frac{\binom{p}{k}\binom{p-k}{k}}{\binom{p}{k}^2} = 1 + o(1)$ by Stirling's Theorem since $k << p$.

Now consider any $k > r \geq 1$. Then one has that the number of tuples for which $|m_1 \cap m_2| = r$ and $\xi_1 = \xi_2$ equals $2\binom{p}{r}\binom{p-r}{k-r}\binom{p-k}{k-r}$ and the number of tuples for which $|m_1 \cap m_2| = r$ and $\xi_1 = -\xi_2$ also equals $2\binom{p}{r}\binom{p-r}{k-r}\binom{p-k}{k-r}$. Hence the total sum can be bounded by $2^r \frac{1}{4} \frac{1}{\binom{p}{k}^2} 2\binom{p}{r}\binom{p-r}{k-r}\binom{p-k}{k-r}\{[\frac{1+e^{2A}}{(1+e^A)^2}]^r + [\frac{2e^A}{(1+e^A)^2}]^r\} \leq 2^r \frac{1}{\binom{p}{k}^2}\binom{p}{r}\binom{p-r}{k-r}\binom{p-k}{k-r}$ because $[\frac{1+e^{2A}}{(1+e^A)^2}]^r + [\frac{2e^A}{(1+e^A)^2}]^r \leq 2$. Hence,

$$
\begin{aligned}
\mathbb{E}_0(L_\pi^2) &\le \frac{\binom{p}{k}\binom{p-k}{k}}{\binom{p}{k}^2} + \sum_{r=1}^{k} \frac{2^r \binom{p}{r}\binom{p-r}{k-r}\binom{p-k}{k-r}}{\binom{p}{k}^2} \\
&= \frac{\binom{p}{k}\binom{p-k}{k}}{\binom{p}{k}^2} + \frac{2^k}{\binom{p}{k}} + \sum_{r=1}^{k-1} 2^r \frac{(p-k)\cdots(p-2k+r+1)}{p\cdots(p-k+1)} \frac{k!k!}{r!(k-r)!(k-r)!} \\
&\le \frac{\binom{p-k}{k}}{\binom{p}{k}} + \frac{2^k}{\binom{p}{k}} + \sum_{r=1}^{k-1} 2^r \frac{(p-k+1)^{k-r}}{(p-k+1)^k \frac{[k\cdots(k-r+1)]^2}{r!}} \\
&\le \frac{\binom{p-k}{k}}{\binom{p}{k}} + \frac{2^k}{\binom{p}{k}} + \sum_{r=1}^{k-1} 2^r \frac{1}{(p-k+1)^r} \frac{k^{2r}}{r!} \le \frac{\binom{p-k}{k}}{\binom{p}{k}} + \frac{2^k}{\binom{p}{k}} + \sum_{r=1}^{k-1} \left(\frac{2k^2}{(p-k+1)}\right)^r \frac{1}{r^r e^{-r}} \\
&\le \frac{\binom{p-k}{k}}{\binom{p}{k}} + \frac{2^k}{\binom{p}{k}} + \sum_{r=1}^{k-1} \left(\frac{2ek^2}{(p-k+1)}\right)^r = \frac{\binom{p-k}{k}}{\binom{p}{k}} + \frac{2^k}{\binom{p}{k}} + \frac{1-(\frac{2ek^2}{(p-k+1)})^{k-1}}{1-\frac{2ek^2}{(p-k+1)}} - 1
\end{aligned}
$$

The last step holds because $k^2 \ll p$ since $\alpha > 1/2$. For $r = k$ we have the factor $\binom{p}{r}\binom{p-r}{k-r}\binom{p-k}{k-r}$ replaced by $\binom{p}{k}$. Now since $\frac{2^k}{\binom{p}{k}} \le (\frac{2k}{p-k+1})^k = o(1)$ and $\frac{1-(\frac{2ek^2}{(p-k+1)})^{k-1}}{1-\frac{2ek^2}{(p-k+1)}} = 1+o(1)$ we have that $\mathbb{E}_0(L_\pi^2) \le 1 + o(1)$.

□

Proof of Theorem 2.6. We first present the proof of the lower bound.

We will estimate the second moment of the likelihood ratio as follows.

$$
\begin{aligned}
\mathbb{E}_0(L_\pi^2) &= 2^{-2k}\binom{p}{k}^{-2} \sum_{m_1,m_2,\xi_1,\xi_2} \left(\frac{1+4\Delta^2}{1-4\Delta^2}\right)^{r|m_3|} (1-4\Delta^2)^{r|m_1\cap m_2|} \\
&= \mathbb{E}_{|m_3|,|m_1\cap m_2|}\left[\left(\frac{1+4\Delta^2}{1-4\Delta^2}\right)^{r|m_3|} (1-4\Delta^2)^{r|m_1\cap m_2|}\right]
\end{aligned}
$$

where $m_3 = \{j \in m_1 \cap m_2 : \xi_1^j = \xi_2^j\}$. Now given $|m_1 \cap m_2|$, $|m_3| \sim \text{Bin}(|m_1 \cap m_2|, \frac{1}{2})$. Hence

$$
\begin{aligned}
\mathbb{E}_0(L_\pi^2) &= \mathbb{E}_{|m_1\cap m_2|}\left[\left(\frac{1}{2} + \frac{1}{2}\left(\frac{1+4\Delta^2}{1-4\Delta^2}\right)^r\right)^{|m_1\cap m_2|} (1-4\Delta^2)^{r|m_1\cap m_2|}\right] \qquad \text{(B.4)} \\
&= \mathbb{E}_{|m_1\cap m_2|}\left[\left(\frac{1}{2}\right)^{|m_1\cap m_2|} \left((1+4\Delta^2)^r + (1-4\Delta^2)^r\right)^{|m_1\cap m_2|}\right] \\
&= \mathbb{E}_Z\left[\left(\frac{1}{2}\right)^Z (a^r+b^r)^Z\right] = \mathbb{E}_Z\left[2^{(r-1)Z}(a_1^r+b_1^r)^Z\right]
\end{aligned}
$$

where $Z \sim \text{Hypergeometric}(p,k,k)$ and $a = (1+4\Delta^2)^r$, $b = (1-4\Delta^2)^r$ and $(a_1,b_1) =$

$(a/2, b/2)$. Thus $a_1 + b_1 = 1$ and hence $(a_1^r + b_1^r)2^{r-1} \geq 1$. Now since $Z \sim \text{Hypergeometric}(p, k, k)$, Z is stochastically smaller than W where $W \sim \text{Bin}(k, \frac{k}{p-k})$. Hence

$$\begin{aligned}
\mathbb{E}_0(L_\pi^2) &= \mathbb{E}_Z\left[2^{(r-1)Z}(a_1^r + b_1^r)^Z\right] \\
&\leq \mathbb{E}_W\left[2^{(r-1)W}(a_1^r + b_1^r)^W\right] \\
&= \left[1 + \frac{\frac{k^2}{p-k}(2^{r-1}(a_1^r + b_1^r) - 1)}{k}\right]^k
\end{aligned}$$

We complete our proof by showing that $\frac{k^2}{p-k}(2^{r-1}(a_1^r + b_1^r) - 1) \to 0$ when $\Delta \ll \sqrt{\frac{p^{\frac{1}{2}}}{kr}}$ and hence rendering all tests asymptotically powerless. To this end, note that by Taylor series expansion up to 4^{th} order around 0 and analyzing the remainder, we have

$$\begin{aligned}
\frac{k^2}{p-k}(2^{r-1}(a_1^r + b_1^r) - 1) &= \frac{k^2}{p-k}\left(192\frac{\Delta^4}{4!}r(r-1) + O(\Delta^4 r^2)\right) \\
&= O\left(\frac{k^2 r^2 \Delta^4}{p}\right) \to 0
\end{aligned}$$

where the last line holds since $\Delta \ll \sqrt{\frac{p^{\frac{1}{2}}}{kr}}$. This completes the proof of the lower bound for problem 2.13. The proof of lower bound in 2.4 follows by noting that $\theta(A) = \frac{1}{2} + \Delta$ and the fact that $\theta \in BC^1(0)$.

Now we prove the upper bound. Recall T_{GLRT} from (2.9). Once again we will provide proof for problem 2.13. The proof of lower bound in problem 2.4 follows by noting that $\theta(A) = \frac{1}{2} + \Delta$ and the fact that $\theta \in BC^1(0)$.

We will show that if $t_p \to \infty$ at a sufficiently slow rate, the test is asymptotically powerful. It suffices to show $\sup_{\nu \in \Theta_k^A} \mathbb{P}_\nu(\frac{\mathsf{T}_{\text{GLRT}} - p}{\sqrt{2p}} \leq t_p) \to 0$. We will show that $\sup_{\nu \in \Theta_k^A} \frac{\mathbb{E}_\nu(\frac{\mathsf{T}_{\text{GLRT}} - p}{\sqrt{2p}})}{t_p} \to \infty$ and $\frac{\text{Var}_\nu(\frac{\mathsf{T}_{\text{GLRT}} - p}{\sqrt{2p}})}{(\mathbb{E}_\nu(\frac{\mathsf{T}_{\text{GLRT}} - p}{\sqrt{2p}}))^2} \to 0$ when $\frac{A^2 kr}{\sqrt{p}} \to \infty$.

Fix $\nu^* \in \Xi_k^\Delta$. Under the measure $\mathbb{P}_{\nu^*}$, exactly k of the Z_j's are distributed as i.i.d $\text{Bin}(r, \frac{1}{2} + \Delta)$ and the rest of the $p - k$ Z_j's are distributed as i.i.d $\text{Bin}(r, \frac{1}{2})$. Let $O = \{j : \beta_j^* \neq 0\}$. Hence we have, for $j \in O$,

$$\mathbb{E}_{\nu^*}\left[(Z_j - \frac{r}{2})^2\right] = r\left(\frac{1}{4} - \Delta^2\right) + r^2\Delta^2 \,. \tag{B.5}$$

For $j \in O^c$, $\mathbb{E}_{\nu^*}[(Z_j - \frac{r}{2})^2] = \frac{r}{4}$. Hence,

$$\frac{\mathbb{E}_{\nu^*}(\frac{\mathsf{T}_{\mathrm{GLRT}}-p}{\sqrt{2p}})}{t_p} = \frac{\frac{\frac{4}{r}[kr(\frac{1}{2}-\Delta^2)+kr^2\Delta^2+(p-k)\frac{r}{4}]-p}{\sqrt{2p}}}{t_p}$$
$$= \frac{\frac{p+4kr\Delta^2-\frac{k\Delta^2}{4}-p}{\sqrt{2p}}}{t_p} \gtrsim \frac{kr\Delta^2}{t_p\sqrt{p}} \approx \frac{krA^2}{\sqrt{p}t_p} \,. \tag{B.6}$$

Since $\frac{krA^2}{\sqrt{p}} \to \infty$ and t_p can be chosen to grow to ∞ at a sufficiently slow rate, (B.6) goes to infinity.

Now we compute the variance. For $j \in O$,

$$\mathbb{E}_{\nu^*}\left(Z_j - \frac{r}{2}\right)^4 = r\left(\frac{1}{4} - \Delta^2\right)\left[3r\left(\frac{1}{4} - \Delta^2\right) + 6\Delta\left(\frac{1}{2} + \Delta\right) - 8r\Delta^2 + 6r\Delta^2\right] + r^4\Delta^4 \,.$$

Using the above and (B.5), a straightforward calculation yields that

$$\sum_{j\in O} \mathrm{Var}_{\nu^*}\left(\frac{4(Z_j - \frac{r}{2})^2}{r}\right) = 16k\left(\frac{1}{4} - \Delta^2\right)\left[2\left(\frac{1}{4} - \Delta^2\right) + 4r\Delta^2 + 6\Delta\left(\frac{1}{2} + \Delta\right) - 8\Delta^2\right] \,.$$

Also, by another direct calculation

$$\sum_{j\in O^c} \mathrm{Var}_{\nu^*}\left(\frac{4(Z_j - \frac{r}{2})^2}{r}\right) = 2(p-k)(1 - \frac{1}{r})$$

Combining the above two,

$$\mathrm{Var}_{\nu^*}\left(\frac{\mathsf{T}_{\mathrm{GLRT}} - p}{\sqrt{2p}}\right) = \left[16k\left(\frac{1}{4} - \Delta^2\right)\left\{2\left(\frac{1}{4} - \Delta^2\right) + 4r\Delta^2 + 6\Delta\left(\frac{1}{2} + \Delta\right) - 8\Delta^2\right\}\right]/2p$$
$$+ 2(p-k)(1 - \frac{1}{r})/2p$$
$$\leq \frac{4p + 32kr\Delta^2}{2p} = \frac{2p + 16kr\Delta^2}{p} \,.$$

Also $\left(\mathbb{E}_{\nu^*}(\frac{\mathsf{T}_{\mathrm{GLRT}}-p}{\sqrt{2p}})\right)^2 \geq \frac{kr\Delta^2}{4p}$. Hence,

$$\frac{\mathrm{Var}_{\nu^*}(\frac{\mathsf{T}_{\mathrm{GLRT}}-p}{\sqrt{2p}})}{(\mathbb{E}_{\nu^*}(\frac{\mathsf{T}_{\mathrm{GLRT}}-p}{\sqrt{2p}}))^2} \leq 4\frac{2p + 16kr\Delta^2}{kr\Delta^2} \to 0$$

since $k^2r^2\Delta^4 \gg p$.

Now note that if ν^* had k_1 elements which are greater than or equal to A and k_2 elements less than equal to $-A$, then a similar calculation yields $\mathrm{Var}_{\nu^*}(\frac{\mathsf{T}_{\mathrm{GLRT}}-p}{\sqrt{2p}}) \leq \frac{2p+16kr\Delta^2}{p}$ and

$(\mathbb{E}_{\nu^*}(\frac{\mathsf{T}_{\mathrm{GLRT}}-p}{\sqrt{2p}}))^2 \geq \frac{kr\Delta^2}{4p}$ where $k = k_1 + k_2$ equals the number of nonzero coefficients in β^*. Hence we have $\max_{\nu\in\Xi_k^\Delta}[\mathbb{P}_\nu(\mathsf{T}_{\mathrm{GLRT}} \leq t_p)] \to 0$ when $\alpha \leq \frac{1}{2}$ and $\frac{\Delta^2 kr}{\sqrt{p}} \to \infty$. This proves the GLRT is asymptotically powerful.

□

Proof of Theorem 2.7. We will provide an argument for problem 2.13. The proof for problem 2.4 follows from noting that $\theta(A) = \frac{1}{2} + \Delta$.

We will estimate the second moment of the likelihood ratio similar to before. Following the same line of arguments as in proof of Theorem 2.6, we note that

$$\begin{aligned}\mathbb{E}_0(L_\pi^2) &= \mathbb{E}_Z\left[2^{(r-1)Z}\{a_1^r + b_1^r\}^Z\right] \text{ where } Z \sim \mathrm{Hypergeometric}(p,k,k)\\ &\leq \left[1 + \frac{\frac{k^2}{p-k}(2^{r-1}(a_1^r+b_1^r)-1)}{k}\right]^k .\end{aligned}$$

Now $\alpha > \frac{1}{2}$ implies that $\frac{k^2}{p-k} \to 0$. Also the quantity $\frac{k^2}{p-k}(2^{r-1}(a_1^r+b_1^r)-1) = O(k^2 2^r/p)$. Hence if $r \ll \frac{log(p)}{log(2)}$, we have that $\mathbb{E}_0(L_\pi^2) \to 1$ and thus all tests are asymptotically powerless.

□

Proof of Theorem 2.8. We will provide proof for the lower bound in problem 2.4 where $\theta \in BC^2(0)$. Using Remark 2.1, the proof also holds for problem 2.13. Since directly bounding $\mathbb{E}_0(L_\pi^2)$ yields trivial bounds we invoke a truncation trick which breaks down the analysis into parts related to extreme tails and non-extreme tails of the Z-statistics. In particular, define the interval

$$H_p = \left(\frac{r}{2} - \sqrt{2log(p)}\sqrt{\frac{r}{4}}, \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right) . \tag{B.7}$$

and put

$$D = \{Z_l \in H_p,\, l = 1,\ldots,p\},\;\; Z_l = \sum_{s=1}^{r} y_{(l-1)r+s},\; l = 1,\ldots,p. \tag{B.8}$$

By Hölder's inequality it can be shown that for proving a lower bound it suffices to prove,

$$\mathbb{E}_0(L_\pi\, \mathbf{I}_{D^c}) = o(1), \qquad \mathbb{E}_0(L_\pi^2\, \mathbf{I}_D) = 1 + o(1) . \tag{B.9}$$

We first prove the first inequality of (B.9). Since $\mathbf{I}_{D^c} \leq \sum_{l=0}^{p-1} \mathbf{I}_{(Z_{l+1} \in H_p^c)}$ and

$$L_\pi = 2^n \int \prod_{j=1}^{p} \left\{ \frac{\theta(\beta_j)}{\theta(-\beta_j)} \right\}^{Z_j} \{\theta(-\beta_j)\}^r d\pi(\boldsymbol{\beta})$$

we have

$$L_\pi \mathbf{I}_{D^c} \leq 2^n \int \sum_{l=1}^{p} \prod_{j=1}^{p} \left\{ \frac{\theta(\beta_j)}{\theta(-\beta_j)} \right\}^{Z_j} \{\theta(-\beta_j)\}^r \mathbf{I}(Z_l \in H_p^c) d\pi(\boldsymbol{\beta})$$

Hence

$$\begin{aligned}
&\mathbb{E}_0(L_\pi \mathbf{I}_{D^c}) \\
&\leq 2^n \int \sum_{l=1}^{p} \mathbb{E}_0 \left[\prod_{j=1}^{p} \left\{ \frac{\theta(\beta_j)}{\theta(-\beta_j)} \right\}^{Z_j} \{\theta(-\beta_j)\}^r \mathbf{I}(Z_l \in H_p^c) \right] d\pi(\boldsymbol{\beta}) \\
&= 2^n \int \sum_{l=1}^{p} \mathbb{E}_0 \left[\prod_{j \neq l}^{p} \left\{ \frac{\theta(\beta_j)}{\theta(-\beta_j)} \right\}^{Z_j} \{\theta(-\beta_j)\}^r \right] \mathbb{E}_0 \left[\left\{ \frac{\theta(\beta_l)}{\theta(-\beta_l)} \right\}^{Z_l} \{\theta(-\beta_l)\}^r \mathbf{I}(Z_l \in H_p^c) \right] d\pi(\boldsymbol{\beta}) \\
&= 2^n \int \sum_{l=1}^{p} \left[\prod_{j \neq l}^{p} \left(\frac{1}{2} \right)^r \left(1 + \frac{\theta(\beta_j)}{\theta(-\beta_j)} \right)^r \{\theta(-\beta_j)\}^r \right] \\
&\qquad \times \mathbb{E}_0 \left[\left\{ \frac{\theta(\beta_l)}{\theta(-\beta_l)} \right\}^{Z_l} \{\theta(-\beta_l)\}^r \mathbf{I}(Z_l \in H_p^c) \right] d\pi(\boldsymbol{\beta}) \\
&= 2^n \int \sum_{l=1}^{p} \left(\frac{1}{2} \right)^{(p-1)r} \mathbb{E}_0 \left[\left\{ \frac{\theta(\beta_l)}{\theta(-\beta_l)} \right\}^{Z_l} \{\theta(-\beta_l)\}^r \mathbf{I}(Z_l \in H_p^c) \right] d\pi(\boldsymbol{\beta}) \\
&= \int \sum_{l=1}^{p} 2^r \mathbb{E}_0 \left[\left\{ \frac{\theta(\beta_l)}{\theta(-\beta_l)} \right\}^{Z_l} \{\theta(-\beta_l)\}^r \mathbf{I}(Z_l \in H_p^c) \right] d\pi(\boldsymbol{\beta})
\end{aligned}$$

Letting $m_1^1 = \{j \in m_1 : \xi_1^j = +1\}$ and $m_1^{-1} = \{j \in m_1 : \xi_1^j = -1\}$, we have

$$
\begin{aligned}
&\mathbb{E}_0(L_\pi \mathbf{I}_{D^c}) \\
&\leq \binom{p}{k}^{-1} 2^{-k} 2^r \sum_{m_1,\xi_1} \Big[\sum_{j \in m_1^1} \mathbb{E}_0\Big(\Big\{ \frac{\theta(A)}{\theta(-A)} \Big\}^{Z_j} \Big\{ \theta(-A) \Big\}^r \mathbf{I}(Z_j \in H_p^c) \Big) \\
&\quad + \sum_{j \in m_1^{-1}} \mathbb{E}_0\Big(\Big\{ \frac{\theta(-A)}{\theta(A)} \Big\}^{Z_j} \{\theta(A)\}^r \mathbf{I}(Z_j \in H_p^c) \Big) \\
&\quad + \sum_{j \in m_1^c} \mathbb{E}_0\Big(\Big\{ \frac{\theta(0)}{\theta(-0)} \Big\}^{Z_j} \{\theta(-0)\}^r \mathbf{I}(Z_j \in H_p^c) \Big) \Big] \\
&= \binom{p}{k}^{-1} 2^{-k} 2^r \sum_{m_1,\xi_1} \Big[\sum_{j \in m_1^1} \mathbb{E}_0\Big(\Big\{ \frac{\theta(A)}{\theta(-A)} \Big\}^{Z_j} \{\theta(-A)\}^r \mathbf{I}(Z_j \in H_p^c) \Big) \\
&\quad + \sum_{j \in m_1^{-1}} \mathbb{E}_0\Big(\Big\{ \frac{\theta(A)}{\theta(-A)} \Big\}^{r - Z_j} \{\theta(-A)\}^r \mathbf{I}(Z_j \in H_p^c) \Big) + \sum_{j \in m_1^c} \Big(\frac{1}{2} \Big)^r \mathbb{P}(Z_j \in H_p^c) \Big] \\
&= k \{2\theta(-A)\}^r \mathbb{E}_0\Big(\Big\{ \frac{\theta(A)}{\theta(-A)} \Big\}^{Z_1} \mathbf{I}(Z_1 \in H_p^c) \Big) + (p-k) \mathbb{P}(Z_1 \in H_p^c) \qquad \text{(B.10)}
\end{aligned}
$$

where we have used the fact that $r - Z_l \stackrel{d}{=} Z_l$ and that the set D in (B.8) is symmetric in Z_l and $r - Z_l$.

Now by Lemma B.2 we have that $(p-k)\mathbb{P}(Z_1 \in H_p^c) = o(1)$ since $r \gg log(p)$. To see this we put $n = r$ and $k' = \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}$ in Lemma B.2 to obtain $\epsilon = \frac{2\sqrt{\frac{r}{4}}\sqrt{2log(p)} - 1}{r-1} = o(1)$ since $r \gg log(p)$ and also $\epsilon\sqrt{r} \to \infty$. This implies

$$
\begin{aligned}
\mathbb{P}\Big(Z_l > \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}} \Big) &= \overline{\Phi}(\epsilon\sqrt{n}) e^{-(r-1)((1+\epsilon)log(1+\epsilon) + (1-\epsilon)log(1-\epsilon) - \epsilon^2)} \\
&\leq \frac{e^{-\frac{4.\frac{1}{4}.2log(p)}{2}}}{\epsilon\sqrt{r}} e^{r\epsilon^2 - r(1+\epsilon)log(1+\epsilon)} \text{ using Lemma A.1} \\
&\leq \frac{e^{-log(p)}}{\epsilon\sqrt{r}} e^{r\epsilon^2 - r\epsilon} \text{ since } log(1+\epsilon) \geq \frac{\epsilon}{(1+\epsilon)} \\
&\ll \frac{1}{p\epsilon\sqrt{r}}. \qquad \text{(B.11)}
\end{aligned}
$$

Hence $(p-k)\mathbb{P}(Z_1 \in H_p^c) = o(1)$ as needed. Next we need to control $k\{2\theta(-A)\}^r \mathbb{E}_0 \{\frac{\theta(A)}{\theta(-A)}\}^{Z_1} \mathbf{I}(Z_1 \in H_p^c)$. To this end note that

$$
\Big\{ \frac{\theta(A)}{\theta(-A)} \Big\}^{Z_1} = e^{Z_1 log(\frac{\theta(A)}{\theta(-A)})} = e^{(2\frac{\theta'(0)}{\theta(0)} A + o(A^2)) Z_1} = e^{(4\theta'(0)A + o(A^2))Z_1}.
$$

Hence by Hölder's Inequality for any $f > 1$ and complementary $g > 1$ such that $\frac{1}{f} + \frac{1}{g} = 1$, one has

$$k\{2\theta(-A)\}^r \mathbb{E}_0\Big(\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_1}\mathbf{I}(Z_1 \in H_p^c)\Big) \leq \{k^f\{2\theta(-A)\}^{rf}\mathbb{E}_0[e^{4\theta'(0)AfZ_1}\mathbf{I}(Z_1 \in H_p^c)]\}^{1/f} \times \{\mathbb{E}_0[e^{g\epsilon Z_1}]\}^{1/g} \tag{B.12}$$

where $\epsilon = o(A^2)$. Our next task is hence to control $k^f\{2\theta(-A)\}^{rf}\mathbb{E}_0[e^{4\theta'(0)AfZ_1}\mathbf{I}(Z_1 \in H_p^c)]$ for an appropriately chosen $f > 1$ and then subsequently bound $\{\mathbb{E}_0[e^{g\epsilon Z_1}]\}^{1/g}$ for the corresponding $g > 1$. We first analyze $\mathbb{E}_0[e^{4\theta'(0)AfZ_1}\mathbf{I}(Z_1 \in H_p^c)]$ for arbitrary $f > 1$ and we will make the choice of the pair (f, g) clear later:

$$\mathbb{E}_0[e^{4\theta'(0)AfZ_1}\mathbf{I}(Z_1 \in H_p^c)] = \mathbb{E}_0\left[e^{4\theta'(0)AfZ_1}\mathbf{I}\left(Z_1 > \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right)\right] + \mathbb{E}_0\left[e^{4\theta'(0)AfZ_1}\mathbf{I}\left(Z_1 < \frac{r}{2} - \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right)\right] := I_1 + I_2.$$

We will analyze I_1 in detail; the analysis of I_2 is very similar and is omitted. Since

$$I_1 = e^{4\theta'(0)f\frac{Ar}{2}}\mathbb{E}_0\left[e^{4\theta'(0)f\frac{A^*}{2}\frac{Z_1-\frac{r}{2}}{\sqrt{\frac{r}{4}}}}\mathbf{I}\left(Z_1 > \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right)\right]$$

where $A^* = A\sqrt{r}$, we will first control $\mathbb{E}_0\left[e^{4\theta'(0)f\frac{A^*}{2}\frac{Z_1-\frac{r}{2}}{\sqrt{\frac{r}{4}}}}\mathbf{I}\left(Z_1 > \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right)\right] = I_1'$ (say). Denoting $\frac{Z_1-\frac{r}{2}}{\sqrt{\frac{r}{4}}}$ by W_r, by the Komlos-Major-Tusnady strong embedding theorem (Komlós et al., 1975), there exists a version of standard Brownian Motion B_r on the same probability space as W_r such that

$$\mathbb{P}(|W_r - B_r| \geq Clog(r) + s) \leq Ke^{-\lambda s} \tag{B.13}$$

where C, K, λ do not depend on r. For notational conveneience we will take $C = 1$ w.l.o.g.

Let $x > 0$ which we will choose appropriately later. Hence

$$\begin{aligned}
I_1' &= \mathbb{E}_0\left[e^{4\theta^{'}(0)f\frac{A^*}{2\sqrt{r}}W_r}\mathbf{I}\left(Z_1 > \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right)\right] \\
&= \mathbb{E}_0\Big[e^{4\theta^{'}(0)f\frac{A}{2}(W_r)}\mathbf{I}(W_r > \sqrt{2log(p)}\sqrt{r}) \\
&\qquad \times \{\mathbf{I}(|W_r - B_r| \leq log(r) + x) + \mathbf{I}(|W_r - B_r| > log(r) + x)\}\Big] \\
&:= I_{11} + I_{12}.
\end{aligned}$$

Hence we will need to control both $k^f\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}}I_{11}$ and $k^f\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}}I_{12}$. Now

$$\begin{aligned}
&I_{11} \\
&= \mathbb{E}_0[e^{4\theta^{'}(0)f\frac{A}{2}(W_r)}\mathbf{I}(W_r > \sqrt{2log(p)}\sqrt{r})\mathbf{I}(|W_r - B_r| \leq log(r) + x)] \\
&\leq e^{4\theta^{'}(0)f\frac{A}{2}(log(r)+x)}\mathbb{E}_0[e^{4\theta^{'}(0)f\frac{A}{2}B_r}\mathbf{I}(B_r > \sqrt{2log(p)}\sqrt{r} - (log(r) + x))] \\
&= e^{4\theta^{'}(0)f\frac{A}{2}(log(r)+x)}\mathbb{E}_0[e^{4\theta^{'}(0)f\frac{A^*}{2}\frac{B_r}{\sqrt{r}}}\mathbf{I}(\frac{B_r}{\sqrt{r}} > \sqrt{2log(p)} - \frac{(log(r) + x)}{\sqrt{r}})] \\
&= e^{4\theta^{'}(0)f\frac{A}{2}(log(r)+x)}\int_{T_p}^{\infty}\frac{e^{4\theta^{'}(0)f\frac{A^*}{2}v - \frac{v^2}{2}}}{\sqrt{2\pi}}dv \text{ where } T_p = \sqrt{2log(p)} - \frac{(log(r) + x)}{\sqrt{r}} \\
&= e^{4\theta^{'}(0)f\frac{A}{2}(log(r)+x) + 2\theta^{'}(0)^2f^2(A^*)^2}\overline{\Phi}(T_p - 2\theta^{'}(0)fA^*) \\
&\leq Ce^{\{4\theta^{'}(0)f\frac{A}{2}(log(r)+x) + 2\theta^{'}(0)^2f^2(A^*)^2 - \frac{T_p^2 - 4\theta^{'}(0)^2(A^*)^2f^2 + 4\theta^{'}(0)A^*fT_p}{2}\}} \text{ if } T_p - 2\theta^{'}(0)fA^* > 1 \\
&= Ce^{\{-log(p)(1 - 4\theta^{'}(0)\sqrt{t}f) - \frac{(log(r)+x)^2}{2r} + \frac{\sqrt{2log(p)}(log(r)+x)}{2\sqrt{r}}\}} \qquad \text{(B.14)}
\end{aligned}$$

Since I_{11} is multiplied outside by $\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}}$ we bound that coefficient as follows:

$$\begin{aligned}
\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}} &= (2\theta(-A)e^{2\theta^{'}(0)A})^{rf} \\
&= e^{rflog(2\theta(-A)e^{2\theta^{'}(0)A})} = e^{rf(log2 + log\theta(_A) + 2\theta^{'}(0)A)} \\
&= e^{rf\{log2 + 2\theta^{'}(0)A + log\theta(0) + \frac{\theta^{'}(0)}{\theta(0)}A(-1) - \frac{1}{2!}\frac{\theta^{''}(0)\theta(0)(-1) - \theta^{'}(0)^2(-1)}{\theta(0)^2}A^2 + o(A^2)\}} \\
&= e^{rf\{log2 + 2\theta^{'}(0)A - log2 - 2\theta^{'}(0)A - 2\theta^{'}(0)^2A^2 + o(A^2)\}} \text{ since } \theta^{''}(0) = 0 \\
&= e^{\{-f4\theta^{'}(0)^2tlog(p) + rf\epsilon^{'}\}} \text{ where } \epsilon' = o(A^2) \qquad \text{(B.15)}
\end{aligned}$$

Finally collecting the terms from (B.14) and (B.15), we bound $k^f\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}}I_{11}$ as

follows:

$$k^f\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}}I_{11} \leq Ce^{-log(p)\{f(1-\alpha-(1-2\theta^{'}(0)\sqrt{t})^2)+(f-1)\}-\frac{(log(r)+x)^2}{2r}+\frac{\sqrt{2log(p)}(log(r)+x)}{2\sqrt{r}}+rf\epsilon'}. \tag{B.16}$$

Now since $t < \rho^*_{\text{binary}}(\alpha)$, $1-\alpha-(1-2\theta^{'}(0)\sqrt{t})^2 < 0$. Hence we can choose $f > 1$ sufficiently close to 1 such that $f(1-\alpha-(1-2\theta^{'}(0)\sqrt{t})^2)+(f-1) < 0$. We note that since $r \gg log(p)$, there exists a sequence $a_{r,p} \to \infty$ such that $r \gg a_{r,p}log(p)$. If we chose $x = a_{r,p}log(p)$ then $T_p - 2\theta^{'}(0)A^*f > 1$ as required for the conclusions to hold since $4\theta^{'}(0)^2t < 1$ and $r \gg a_{r,p}log(p)$. Also again since $r \gg a_{r,p}log(p)$ we have $-log(p)\{f(1-\alpha-(1-2\theta^{'}(0)\sqrt{t})^2)+(f-1)\}-\frac{(log(r)+x)^2}{2r}+\frac{\sqrt{2log(p)}(log(r)+x)}{2\sqrt{r}}+rf\epsilon' \leq -\delta log(p)$ for some $\delta > 0$ for sufficiently large r, p. Hence for such x, we have $k^f\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}}I_{11} \to 0$. In order to bound I_{12} from above we repeatedly apply the Cauchy-Schwarz Inequality and use the fact that $\cosh(s) = 1 + s^2/2 + o(s^2)$ for small s as follows:

$$\begin{aligned}
I_{12} &= \mathbb{E}_0[e^{4\theta^{'}(0)f\frac{A}{2}W_r}\mathbf{I}(W_r > \sqrt{2log(p)}\sqrt{r})\mathbf{I}(|W_r - B_r| > (log(r)+x))] \\
&\leq \left\{\mathbb{E}_0[e^{4\theta^{'}(0)fAW_r}\mathbf{I}(W_r > \sqrt{2log(p)}\sqrt{r})]\mathbb{P}(|W_r - B_r| > (log(r)+x))\right\}^{\frac{1}{2}} \\
&\leq \left\{\mathbb{E}_0[e^{8\theta^{'}(0)fAW_r}]\mathbb{P}(W_r > \sqrt{2log(p)}\sqrt{r})(\mathbb{P}(|W_r - B_r| > (log(r)+x)))^2\right\}^{\frac{1}{4}} \\
&= \left\{(\cosh(8\theta^{'}(0)fA))^r\mathbb{P}(W_r > \sqrt{2log(p)}\sqrt{r})(\mathbb{P}(|W_r - B_r| > (log(r)+x)))^2\right\}^{\frac{1}{4}} \\
&= \left\{e^{rlog(1+32\theta^{'}(0)^2f^2A^2+o(A^2))}\mathbb{P}(W_r > \sqrt{2log(p)}\sqrt{r})(\mathbb{P}(|W_r - B_r| > (log(r)+x)))^2\right\}^{\frac{1}{4}} \\
&\leq \left\{e^{r(32\theta^{'}(0)^2f^2A^2+o(A^2))}\mathbb{P}(W_r > \sqrt{2log(p)}\sqrt{r})(\mathbb{P}(|W_r - B_r| > (log(r)+x)))^2\right\}^{\frac{1}{4}} \\
&\leq C\left\{e^{\{64\theta^{'}(0)^2f^2tlog(p)-log(p)+\frac{(log(p))^2}{r}-\frac{loglog(p)}{2}-2\lambda x+r\epsilon^{''}\}}\right\}^{\frac{1}{4}}
\end{aligned} \tag{B.17}$$

where $\epsilon^{''} = o(A^2)$, the second last line uses the fact that $log(1+x) \leq x$ for $x \geq 0$ and the last line follows from (B.11) and (B.13) for some constant $C > 0$. Recall from (B.15) that $k^f\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}} = e^{\{-4\theta^{'}(0)^2ftlog(p)+rf\epsilon'+(1-\alpha)log(p)\}}$ where $\epsilon' = o(A^2)$. Hence by combining terms from (B.17) and (B.15), we obtain that for a constant K depending on f, t and $\theta^{'}(0)$,

$$k^f\{2\theta(-A)\}^{rf}e^{4\theta^{'}(0)f\frac{Ar}{2}}I_{12} \leq Ce^{Klog(p)-2\lambda x}.$$

Now since $x = a_{r,p}log(p)$ for some $a_{r,p} \to \infty$ such that $r \gg a_{r,p}log(p)$, it follows that

$$k^f \{2\theta(-A)\}^{rf} e^{4\theta'(0)f\frac{Ar}{2}} I_{12} = o(1) \tag{B.18}$$

as required.

Next considering the g-factor from (B.12) we have

$$\begin{aligned}
\{\mathbb{E}_0[e^{g\epsilon Z_1}]\}^{1/g} &= e^{\frac{r}{g}log(\frac{1+e^{g\epsilon}}{2})} \text{ where } \epsilon = o(A^2) \\
&= e^{\frac{r}{g}log(1+\frac{e^{g\epsilon}-1}{2})} = e^{\frac{r}{g}log(1+\frac{g\epsilon+g^2\epsilon^2+o(g^2\epsilon^2)}{2})} \\
&\leq e^{\frac{r}{g}log(1+\frac{(2g+2g^2)\epsilon}{2})} \\
&= e^{r\epsilon O(1)} = e^{o(1)} \to 1
\end{aligned} \tag{B.19}$$

Hence collecting terms from (B.16),(B.18) and (B.19) in (B.12) we finish proving $\mathbb{E}_0(L_\pi \mathbf{I}_{D^c}) = o(1)$ which is the first inequality of (B.9).

Next we prove the second inequality in (B.9). Since definition of D does not depend on $\boldsymbol{\beta}$, it follows that

$$\begin{aligned}
L_\pi^2 \mathbf{I}_D &= (L_\pi \mathbf{I}_D)^2 \\
&= 2^{2n} \iint \prod_{j=1}^{p} \left\{ \frac{\theta(\beta_j)\theta(\beta_j')}{\theta(-\beta_j)\theta(-\beta_j')} \right\}^{Z_j} \left\{\theta(-\beta_j)\theta(-\beta_j')\right\}^r \mathbf{I}(Z_j \in H_p) d\pi(\boldsymbol{\beta}) d\pi(\boldsymbol{\beta})'.
\end{aligned}$$

Hence by Fubini's Theorem and independence of the Z_j's,

$$\begin{aligned}
\mathbb{E}_0(L_\pi^2 \mathbf{I}_D) &= 2^{2n} \iint \prod_{j=1}^{p} \mathbb{E}_0\Big[\left\{ \frac{\theta(\beta_j)\theta(\beta_j')}{\theta(-\beta_j)\theta(-\beta_j')} \right\}^{Z_j} \left\{\theta(-\beta_j)\theta(-\beta_j')\right\}^r \mathbf{I}(Z_j \in H_p) \Big] d\pi(\boldsymbol{\beta}) d\pi(\boldsymbol{\beta}') \\
&= 2^{2n} 2^{-2k} \binom{p}{k}^{-2} \sum_{m_1,m_2,\xi_1,\xi_2} \prod_{j=1}^{p} \mathbb{E}_0\Big[\left\{ \frac{\theta(\beta_j)\theta(\beta_j')}{\theta(-\beta_j)\theta(-\beta_j')} \right\}^{Z_j} \left\{\theta(-\beta_j)\theta(-\beta_j')\right\}^r \mathbf{I}(Z_j \in H_p) \Big]
\end{aligned} \tag{B.20}$$

For any two i.i.d draws (m_1, ξ_1) and $(m_2, , \xi_2)$, set for $j = 1, \ldots, p$

$$T_j = \mathbb{E}_0\Big[\left\{ \frac{\theta(\beta_j)\theta(\beta_j')}{\theta(-\beta_j)\theta(-\beta_j')} \right\}^{Z_j} \left\{\theta(-\beta_j)\theta(-\beta_j')\right\}^r \mathbf{I}(Z_j \in H_p) \Big] .$$

We divide into the following cases. For each $j \in \{1, \ldots, p\}$,

1. $j \in m_1^c \cap m_2^c$: $T_j = \frac{\mathbb{P}(Z_j \in H_p)}{2^r}$.

2. $j \in m_1 \cap m_2^c \cap \{l : \xi_1^l = 1\}$: $T_j = \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_j}\mathbf{I}(Z_j \in H_p)\Big](\frac{\theta(-A)}{2})^r$.

3. $j \in m_1 \cap m_2^c \cap \{l : \xi_1^l = -1\}$:

$$\begin{aligned} T_j &= \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{r-Z_j}\mathbf{I}(Z_j \in H_p)\Big](\frac{\theta(-A)}{2})^r \\ &= \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_j}\mathbf{I}(Z_j \in H_p)\Big](\frac{\theta(-A)}{2})^r \end{aligned}$$

   since $r - Z_j \stackrel{d}{=} Z_j$ and the definition of the set D is also symmetric in Z_j and $r - Z_j$ for all Z_j.

4. $j \in m_1^c \cap m_2 \cap \{l : \xi_2^l = 1\}$: $T_j = \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_j}\mathbf{I}(Z_j \in H_p)\Big](\frac{\theta(-A)}{2})^r$.

5. $j \in m_1^c \cap m_2 \cap \{l : \xi_2^l = -1\}$: $T_j = \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_j}\mathbf{I}(Z_j \in H_p)\Big](\frac{\theta(-A)}{2})^r$ by the symmetry argument made in case (3) above.

6. $j \in m_3 \cap \{l : \xi_1^l = \xi_2^l = 1\}$: $T_j = \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_j}\mathbf{I}(Z_j \in H_p)\Big](\theta(-A))^{2r}$.

7. $j \in m_3 \cap \{j : \xi_1^l = \xi_2^l = -1\}$: $T_j = \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_j}\mathbf{I}(Z_j \in H_p)\Big](\theta(-A))^{2r}$ again by the symmetry argument.

8. $j \in m_4$: $T_j = \{\theta(A)\theta(-A)\}^r\mathbb{P}(Z_j \in H_p)$.

Grouping the terms in (B.20) by the above cases and collecting terms,

$$
\begin{aligned}
&\mathbb{E}_0(L_\pi^2 I(D)) \\
&= \frac{2^{2n-2k}}{\binom{p}{k}^2} \sum_{\substack{m_1,m_2 \\ \xi_1,\xi_2}} \Big[\prod_{j\in m_1^c\cap m_2^c} \frac{\mathbb{P}(Z_j\in H_p)}{2^r} \prod_{j\in m_1^c\Delta m_2^c} \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_j}\mathbf{I}(Z_j\in H_p)\Big]\Big(\frac{\theta(-A)}{2}\Big)^r \\
&\quad \times \prod_{j\in m_3} \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_j}\mathbf{I}(Z_j\in H_p)\Big](\theta(-A))^{2r} \prod_{j\in m_4}\{\theta(A)\theta(-A)\}^r\mathbb{P}(Z_j\in H_p)\Big] \\
&= \frac{2^{2n-2k}}{\binom{p}{k}^2} \sum_{\substack{m_1,m_2 \\ \xi_1,\xi_2}} \Big[\prod_{j\in m_1^c\cap m_2^c} \frac{\mathbb{P}(Z_1\in H_p)}{2^r} \prod_{j\in m_1^c\Delta m_2^c} \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_1}\mathbf{I}(Z_1\in H_p)\Big](\frac{\theta(-A)}{2})^r \\
&\quad \times \prod_{j\in m_3} \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1\in H_p)\Big](\theta(-A))^{2r} \prod_{j\in m_4}\{\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p)\Big] \\
&= \frac{2^{2n-2k}}{\binom{p}{k}^2} \sum_{\substack{m_1,m_2 \\ \xi_1,\xi_2}} \Big[\Big(\frac{\mathbb{P}(Z_1\in H_p)}{2^r}\Big)^{|m_1^c\cap m_2^c|}\Big(\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_1}\mathbf{I}(Z_1\in H_p)\Big](\frac{\theta(-A)}{2})^r\Big)^{|m_1\Delta m_2|} \\
&\quad \times \Big(\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1\in H_p)\Big](\theta(-A))^{2r}\Big)^{|m_3|}\Big(\{\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p)\Big)^{|m_4|}\Big] \\
&= \frac{1}{2^{2k}\binom{p}{k}^2} \sum_{\substack{m_1,m_2 \\ \xi_1,\xi_2}} \Big[\Big(\mathbb{P}(Z_1\in H_p)\Big)^{|m_1^c\cap m_2^c|}\Big(\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{Z_1}\mathbf{I}(Z_1\in H_p)\Big](2\theta(-A))^r\Big)^{|m_1\Delta m_2|} \\
&\quad \times \Big(\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1\in H_p)\Big](2\theta(-A))^{2r}\Big)^{|m_3|}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p)\Big)^{|m_4|}\Big] \\
&\le \frac{1}{2^{2k}\binom{p}{k}^2} \sum_{\substack{m_1,m_2 \\ \xi_1,\xi_2}} \Big[\Big(\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1\in H_p)\Big](2\theta(-A))^{2r}\Big)^{|m_3|} \\
&\quad \times \Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p)\Big)^{|m_4|}\Big] \\
&= \frac{1}{2^{2k}\binom{p}{k}^2} \sum_{\substack{m_1,m_2 \\ \xi_1,\xi_2}} \Big[\Big\{\frac{\Big(\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1\in H_p)\Big](2\theta(-A))^{2r}\Big)}{\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p)\Big)}\Big\}^{|m_3|} \\
&\quad \times \Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p)\Big)^{|m_1\cap m_2|}\Big] \\
&= \frac{1}{\binom{p}{k}^2} \sum_{m_1,m_2} \Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p) \\
&\quad + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1\in H_p)\Big](2\theta(-A))^{2r}\Big)^{|m_1\cap m_2|}\Big] \\
&= \mathbb{E}_W\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1\in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1\in H_p)\Big](2\theta(-A))^{2r}\Big)\Big]^W
\end{aligned}
$$

where $W \sim \text{Hypergeometric}(p, k, k)$. Now we observe that by Lemma B.1,

$$\mathbb{E}_W\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big]^W$$
$$\leq \mathbb{E}_U\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big]^U$$

where $U \sim \text{Bin}(k, \frac{k}{p-k})$, provided the following holds:

$$\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big] \geq 1 \ . \tag{B.21}$$

Hence under the inequality (B.21) we have

$$\mathbb{E}_0(L_\pi^2 I(D))$$
$$\leq \mathbb{E}_U\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big]^U$$
$$= \Big\{1 + \frac{k}{p-k}\Big(\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p)$$
$$+ \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big] - 1\Big)\Big\}$$

Hence in order to prove the second inequality of (B.9) it suffices to verify the inequality (B.21) and prove

$$\frac{k^2}{p}\Big(\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big] - 1\Big) = o(1) \ . \tag{B.22}$$

We first verify (B.22). We note that

$$\frac{k^2}{p}\Big(\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big] - 1\Big)$$
$$:= E_1 + E_2 + E_3 \ . \tag{B.23}$$

where

$$E_1 = \frac{k^2}{2p}\{4\theta(A)\theta(-A)\}^r\mathbb{P}(Z_1 \in H_p),$$
$$E_2 = \frac{k^2}{2p}\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r}$$

and

$$E_3 = \frac{k^2}{p}.$$

Since $\alpha > \frac{1}{2}$, trivially $E_3 = o(1)$. Hence it suffices to prove that $E_1 = o(1)$ and $E_2 = o(1)$. To this end, first note that

$$\begin{aligned} E_1 &= \frac{k^2}{2p}\{4\theta(A)\theta(-A)\}^r \mathbb{P}(Z_1 \in H_p) \leq \frac{k^2}{p}\{4\theta(A)\theta(-A)\}^r \\ &= \frac{k^2}{p} e^{rlog(4\theta(A)\theta(-A))} = \frac{k^2}{p} e^{r(2\theta^{'}(0)A - 2\theta^{'}(0)^2A^2 - 2\theta^{'}(0)A - 2\theta^{'}(0)^2A^2 + o(A^2))} \\ &= \frac{k^2}{p} e^{r(-4\theta^{'}(0)^2A^2 + o(A^2))} = o(1) \end{aligned} \tag{B.24}$$

as required. Next we control E_2 as follows:

$$\begin{aligned} E_2 &= \frac{k^2}{2p}\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \\ &\leq \frac{k^2}{p}\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \\ &= \frac{k^2}{p}\mathbb{E}_0\Big[e^{2Z_1 \log\{\frac{\theta(A)}{\theta(-A)}\}}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \\ &= \mathbb{E}_0\Big[e^{2Z_1(4\theta^{'}(0)A + \epsilon)}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \text{ where } \epsilon = o(A^2) \\ &\leq \Big\{\mathbb{E}_0\Big[e^{8\theta^{'}(0)AZ_1}\mathbf{I}(Z_1 \in H_p)(2\theta(-A))^{2r}\frac{k^2}{p}\Big]^f\Big\}^{1/f}\Big\{\mathbb{E}_0\Big[e^{2g\epsilon Z_1}\Big]\Big\}^{1/g}. \end{aligned} \tag{B.25}$$

where the last line is by Hölder's Inequality for any $f > 1$ and complementary $g > 1$ such that $\frac{1}{f} + \frac{1}{g} = 1$. Our next task is hence to control $\mathbb{E}_0\Big[e^{8\theta^{'}(0)AZ_1}\mathbf{I}(Z_1 \in H_p)(2\theta(-A))^{2r}\frac{k^2}{p}\Big]^f$ for an appropriately chosen $f > 1$ and then subsequently bound $\Big\{\mathbb{E}_0\Big[e^{2g\epsilon Z_1}\Big]\Big\}^{1/g}$ for the corresponding $g > 1$. We first analyze $\mathbb{E}_0\Big[e^{8\theta^{'}(0)AZ_1}\mathbf{I}(Z_1 \in H_p)(2\theta(-A))^{2r}\frac{k^2}{p}\Big]^f$ for arbitrary $f > 1$ and we will make the choice of the pair (f, g) clear later. To that end, we have

$$\begin{aligned} &\mathbb{E}_0\Big[e^{8\theta^{'}(0)AfZ_1}\mathbf{I}(Z_1 \in H_p)(2\theta(-A))^{2rf}\Big] \\ &= \mathbb{E}_0\left[e^{8\theta^{'}(0)AfZ_1}(2\theta(-A))^{2rf}\left\{\mathbf{I}\left(Z_1 \leq \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right) + \mathbf{I}\left(Z_1 \geq \frac{r}{2} - \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right)\right\}\right] \\ &:= I_1 + I_2 - I_3 \end{aligned} \tag{B.26}$$

where $I_1 = \mathbb{E}_0\Big[e^{8\theta^{'}(0)AfZ_1}(2\theta(-A))^{2rf}\{\mathbf{I}(Z_1 \leq \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}})\}\Big]$ and $I_2 = \mathbb{E}_0\Big[e^{8\theta^{'}(0)AfZ_1}(2\theta(-A))^{2rf}\{\mathbf{I}(Z_1 \geq \frac{r}{2} - \sqrt{2log(p)}\sqrt{\frac{r}{4}})\}\Big]$ and I_3 is the remainder. We will analyze I_1 in detail; the analysis of I_2 is very similar and is omitted. The proof of $I_3 = o(1)$ is easier and can be also done following similar techniques and is hence also omitted. Recalling the definition of $W_r := \frac{Z_1 - \frac{r}{2}}{\sqrt{\frac{r}{4}}}$, we have

$$\mathbb{E}_0\left[e^{8\theta^{'}(0)AfZ_1}\mathbf{I}\left(Z_1 \leq \frac{r}{2} + \sqrt{2log(p)}\sqrt{\frac{r}{4}}\right)\right] = e^{4\theta^{'}(0)fAr}\mathbb{E}_0\left[e^{4\theta^{'}(0)fAW_r}\mathbf{I}\left(\frac{W_r}{\sqrt{r}} \leq \sqrt{2log(p)}\right)\right]$$

Arguing similarly as in proof of the first inequality of (B.9), it can be shown that it suffices to analyze $e^{4\theta^{'}(0)fAr}\mathbb{E}_0\Big[e^{4\theta^{'}(0)fAB_r}\{\mathbf{I}(\frac{B_r}{\sqrt{r}} \leq \sqrt{2log(p)})\}\Big]$ where B_r is the version of Brownian Motion on the same probability space as W_r satisfying (B.13). Of course in the proof of the first inequality of (B.9) we went through complete details in choosing an appropriate $x > 0$ which calibrates the degree of approximation between W_r and B_r. However we note that the same choice of x as before goes through and the essence of the proof boils down to controlling $e^{4\theta^{'}(0)fAr}\mathbb{E}_0\Big[e^{4\theta^{'}(0)fAB_r}\{\mathbf{I}(\frac{B_r}{\sqrt{r}} \leq \sqrt{2log(p)})\}\Big]$. Now

$$\begin{aligned}
&\mathbb{E}_0\left[e^{4\theta^{'}(0)fAB_r}\mathbf{I}\left(\frac{B_r}{\sqrt{r}} \leq \sqrt{2log(p)}\right)\right] \\
&= \mathbb{E}_0\left[e^{4\theta^{'}(0)fA^*\frac{B_r}{\sqrt{r}}}\mathbf{I}\left(\frac{B_r}{\sqrt{r}} \leq \sqrt{2log(p)}\right)\right] \\
&= \int_{-\infty}^{\sqrt{2log(p)}} e^{4\theta^{'}(0)fA^*v}\frac{e^{-\frac{v^2}{2}}}{\sqrt{2\pi}}dv \\
&= \int_{-\infty}^{\sqrt{2log(p)}} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(v^2 - 8\theta^{'}(0)fA^*v + 16\theta^{'}(0)^2f^2(A^*)^2)}e^{8\theta^{'}(0)^2f^2(A^*)^2}dv \\
&= \Phi(\sqrt{2log(p)} - 4\theta^{'}(0)fA^*)e^{8\theta^{'}(0)^2f^2(A^*)^2}
\end{aligned}$$

Considering the expression for I_1 in (B.26), we have the following:

$$e^{4\theta^{'}(0)fAr}(2\theta(-A))^{2rf} = e^{-4\theta^{'}(0)^2(A^*)^2f + rf\epsilon^{'}} \text{ where } \epsilon^{'} = o(A^2)$$

since $\theta^{''}(0) = 0$. Hence we have, as in the proof of the first inequality of (B.9),

$$\begin{aligned}
I_1 &\lesssim e^{(1-2\alpha)flog(p) + 8\theta^{'}(0)^2(A^*)^2f^2 - 4\theta^{'}(0)(A^*)^2f + rf\epsilon^{'}}\Phi(\sqrt{2log(p)} - 4\theta^{'}(0)fA^*) \\
&= e^{\{(1-2\alpha)f + 16\theta^{'}(0)^2f^2t - 8\theta^{'}(0)ft\}log(p) + rf\epsilon^{'}}\Phi(\sqrt{2log(p)} - 4\theta^{'}(0)fA^*)
\end{aligned}$$

Now the behavior of the bounds on $\Phi(s)$ is different depending on whether $s \geq 0$ or $s < 0$ and we have $\Phi(s) \leq 1$ when $s \geq 0$ and $\Phi(s) < \phi(s)$ if $s < 0$. But $\sqrt{2log(p)} - 4\theta^{'}(0)fA^{*} \lessgtr 0$ accordingly as $t \gtrless \frac{1}{16\theta^{'}(0)^{2}f^{2}}$. Hence we divide our analysis into two parts according to the range of t.

When $t \leq \frac{1}{16\theta^{'}(0)^{2}f^{2}}$, *i.e.*, $\sqrt{2log(p)} - 4\theta^{'}(0)fA^{*} \geq 0$ we have

$$I_1 \lesssim e^{\{(1-2\alpha)f + 16\theta^{'}(0)^{2}f^{2}t - 8\theta^{'}(0)ft\}log(p) + rf\epsilon^{'}}$$

Now the coefficient of $log(p)$ in the above exponent is

$$\begin{aligned} f\left[(1-2\alpha) + 8\theta^{'}(0)^{2}t\,(2f-1)\right] &= 2f\left[\left(\frac{1}{2}-\alpha\right) + 4\theta^{'}(0)^{2}t\,(2f-1)\right] \\ &= 8f\theta^{'}(0)^{2}\left[\frac{\frac{1}{2}-\alpha}{4\theta^{'}(0)^{2}} + t\,(2f-1)\right] \end{aligned}$$

For $\alpha \leq \frac{3}{4}$, since $t < \rho^{*}_{\text{binary}}(\alpha) = \frac{\alpha - \frac{1}{2}}{4\theta^{'}(0)^{2}}$, we have there exists $\delta_1(\alpha, t) > 0$ such that $\frac{\frac{1}{2}-\alpha}{4\theta^{'}(0)^{2}} + t(2f-1) < 0$ whenever $f = 1 + \delta$ with $\delta \leq \delta_1(\alpha, t)$. For $\alpha > \frac{3}{4}$, since $t \leq \frac{1}{16\theta^{'}(0)^{2}f^{2}}$, $\frac{\alpha - \frac{1}{2}}{4\theta^{'}(0)^{2}}$ is monotone increasing in α and $\rho^{*}_{\text{binary}}(\frac{3}{4}) = \frac{\frac{3}{4}-\frac{1}{2}}{4\theta^{'}(0)^{2}} = \frac{1}{16\theta^{'}(0)^{2}f^{2}}$, we have that there exists $\delta_2(\alpha, t) > 0$ such that $\frac{\frac{1}{2}-\alpha}{4\theta^{'}(0)^{2}} + t(2f-1) < 0$ whenever $f = 1 + \delta$ with $\delta \leq \delta_2(\alpha, t)$.

When $t > \frac{1}{16\theta^{'}(0)^{2}f^{2}}$ we have

$$\begin{aligned} I_1 &\lesssim e^{\{(1-2\alpha)f + 16\theta^{'}(0)^{2}f^{2}t - 8\theta^{'}(0)ft\}log(p) + rf\epsilon^{'}}\phi(\sqrt{2log(p)} - 4\theta^{'}(0)fA^{*}) \\ &= e^{flog(p)(1-2\alpha-8\theta^{'}(0)^{2}t-1+8\theta^{'}(0)\sqrt{t}) + log(p)(f-1) + rf\epsilon^{'}} \\ &= e^{flog(p)(1-2\alpha-8\theta^{'}(0)^{2}t-1+8\theta^{'}(0)\sqrt{t}) + log(p)(f-1) + rf\epsilon^{'}} \\ &= e^{flog(p)\{2(1-\alpha)-2(1-2\theta^{'}(0)\sqrt{t})^{2}\} + (f-1)log(p) + rf\epsilon^{'}} \end{aligned}$$

Since $t < \rho^{*}_{\text{binary}}(\alpha)$, $2(1-\alpha) - 2(1-2\theta^{'}(0)\sqrt{t})^{2} < 0$ and hence there exists $\delta_3(\alpha, t) > 0$ such that $f\{\{2(1-\alpha)-2(1-2\theta^{'}(0)\sqrt{t})^{2}\} + (f-1) < 0$ whenever $f = 1+\delta$ with $\delta \leq \delta_3(\alpha, t)$.

Hence choosing $f = 1 + \delta$ with $\delta = \min\{\delta_1(\alpha, t), \delta_2(\alpha, t), \delta_3(\alpha, t)\}$ yields $I_1 = o(1)$ as required. Controlling the corresponding g-factor in (B.25) is similar to that in (B.12) and

can be done along the lines of deriving (B.19).

Next we prove (B.21). We note that it suffices to prove that $\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \to \infty$. As before

$$\begin{aligned}
&\mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \\
&= \mathbb{E}_0\Big[e^{2Z_1 \log\{\frac{\theta(A)}{\theta(-A)}\}}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \\
&= \mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)A2Z_1}\mathbf{I}(Z_1 \in H_p)\Big](2\theta(-A))^{2r} \text{ where } \epsilon A = o(A^2) \\
&= e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}\mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)AW_r}\mathbf{I}(Z_1 \in H_p)\Big].
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)AW_r}\mathbf{I}(Z_1 \in H_p)\Big] \\
&\geq \mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)AW_r}\mathbf{I}(|W_r - B_r| \leq (log(r)+x))\mathbf{I}(-\sqrt{2log(p)}\sqrt{r} \leq W_r \leq \sqrt{2log(p)}\sqrt{r}) \\
&\geq e^{-(4\theta^{'}(0)+\epsilon)(log(r)+x)A}\mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)AB_r}\mathbf{I}(|W_r - B_r| \leq (log(r)+x)) \\
&\qquad \times \mathbf{I}(-\sqrt{2log(p)}\sqrt{r} + (log(r)+x) \leq B_r \leq \sqrt{2log(p)}\sqrt{r} - (log(r)+x))\Big] \\
&= e^{-(4\theta^{'}(0)+\epsilon)(log(r)+x)A}\mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)A^*\frac{B_r}{\sqrt{r}}} \\
&\quad \mathbf{I}(-\sqrt{2log(p)}\sqrt{r} + \frac{(log(r)+x)}{\sqrt{r}} \leq \frac{B_r}{\sqrt{r}} \leq \sqrt{2log(p)} - \frac{(log(r)+x)}{\sqrt{r}})\Big] \\
&- \quad e^{-(4\theta^{'}(0)+\epsilon)(log(r)+x)A}\mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)A^*\frac{B_r}{\sqrt{r}}} \\
&\quad \mathbf{I}(-\sqrt{2log(p)}\sqrt{r} + \frac{(log(r)+x)}{\sqrt{r}} \leq \frac{B_r}{\sqrt{r}} \leq \sqrt{2log(p)} - \frac{(log(r)+x)}{\sqrt{r}}) \\
&\qquad \times \mathbf{I}(|W_r - B_r| > (log(r)+x))\Big] \\
&:= S_1 - S_2 \ .
\end{aligned} \tag{B.27}$$

Hence it is enough to prove that $e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 \to \infty$ and $e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_2 = O(1)$.

Now

$$S_1 := e^{-(4\theta^{'}(0)+\epsilon)(log(r)+x)A}\mathbb{E}_0\Big[e^{(4\theta^{'}(0)+\epsilon)A^*\frac{B_r}{\sqrt{r}}} \mathbf{I}(-\sqrt{2log(p)}\sqrt{r}+\frac{(log(r)+x)}{\sqrt{r}} \leq \frac{B_r}{\sqrt{r}} \leq \sqrt{2log(p)}-\frac{(log(r)+x)}{\sqrt{r}})\Big]$$
$$= e^{-(4\theta^{'}(0)+\epsilon)(log(r)+x)A}e^{\frac{1}{2}(4\theta^{'}(0)+\epsilon)^2(A^*)^2}\Phi(\sqrt{2log(p)}-\frac{(log(r)+x)}{\sqrt{r}}-(4\theta^{'}(0)+\epsilon)A^*) . \tag{B.28}$$

Also

$$e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r} = e^{-4\theta^{'}(0)^2A^2r+ro(A^2)+\epsilon Ar} = e^{-4\theta^{'}(0)^2A^2r+ro(A^2)}. \tag{B.29}$$

Hence by (B.28) and (B.29) we have

$$e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 = e^{\{\frac{1}{2}(4\theta^{'}(0)+\epsilon)^2(A^*)^2-(4\theta^{'}(0)+\epsilon)(log(r)+x)A-4\theta^{'}(0)^2A^2r+ro(A^2)\}} \times \Phi\Big(\sqrt{2log(p)}-\frac{(log(r)+x)}{\sqrt{r}}-(4\theta^{'}(0)+\epsilon)A^*\Big). \tag{B.30}$$

The behavior of the above quantity depends on $\Phi(\eta)$ where $\eta = \sqrt{2log(p)}-\frac{(log(r)+x)}{\sqrt{r}}-(4\theta^{'}(0)+\epsilon)A^*$. Hence we divide our study in the following cases.

First suppose $t \leq \frac{1}{16\theta^{'}(0)^2}$. If $\epsilon = -\delta < 0$, then $\eta \geq \frac{\sqrt{2log(p)}\delta}{4\theta^{'}(0)}-\frac{(log(r)+x)}{\sqrt{r}}$. Hence $\Phi(\eta) \geq \frac{1}{2}+o(1)$. Hence from (B.30) we have

$$e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1$$
$$\geq (\frac{1}{2}+o(1))e^{\{\frac{1}{2}(4\theta^{'}(0)+\epsilon)^2(A^*)^2-(4\theta^{'}(0)+\epsilon)(log(r)+x)Ar-4\theta^{'}(0)^2A^2r+ro(A^2)\}}$$
$$= (\frac{1}{2}+o(1))e^{4\theta^{'}(0)(A^*)^2+\kappa-(4\theta^{'}(0)+\epsilon)(log(r)+x)A+ro(A^2)} \text{ where } |\kappa| \ll log(p)$$
$$= (\frac{1}{2}+o(1))e^{8t\theta^{'}(0)log(p)+\kappa-(4\theta^{'}(0)+\epsilon)(log(r)+x)A+ro(A^2)}. \tag{B.31}$$

Now $(4\theta^{'}(0)+\epsilon)(log(r)+x)A < \frac{5\theta^{'}(0)(log(r)+x)\sqrt{2log(p)}}{\sqrt{r}} \ll log(p)$ if $x = a_{r,p}log(p)$ is

such that $a_{r,p} \to \infty$ ensuring both $r \gg a_{r,p}log(p)$ and $\frac{a_{r,p}log(p)\sqrt{2log(p)}}{\sqrt{r}} \ll log(p)$. Thus $e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 \geq e^{clog(p)}$ for some $c > 0$ and hence diverges.

If $\epsilon > 0$, then $\eta \geq -\frac{\sqrt{2log(p)}\epsilon}{4\theta^{'}(0)} - \frac{(log(r)+x)}{\sqrt{r}}$ and hence $-\eta \leq \frac{\sqrt{2log(p)}\epsilon}{4\theta^{'}(0)} + \frac{(log(r)+x)}{\sqrt{r}} \ll \tau$ for some divergent $\tau \ll \sqrt{log(p)}$. Hence, by Lemma A.1,

$$\begin{aligned}\Phi(\eta) = \overline{\Phi}(-\eta) &\geq (1 - \frac{1}{\tau^2})\frac{\phi(\tau)}{\tau} \\ &\geq \frac{\phi(\tau)}{2\tau} \text{ for sufficiently large } r, p \\ &= \frac{e^{-\tau^2/2}}{\tau\sqrt{2\pi}}.\end{aligned}$$

Hence similar to the calculations in deriving (B.31) we have

$$\begin{aligned}e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 &\geq \frac{e^{-\tau^2/2}}{\tau\sqrt{2\pi}}e^{8t\theta^{'}(0)log(p)+\kappa-(4\theta^{'}(0)+\epsilon)(log(r)+x)A+ro(A^2)} \\ &\geq \frac{e^{-\tau^2/2}}{\tau\sqrt{2\pi}}e^{clog(p)} \text{ for some } c > 0 \\ &= \frac{e^{clog(p)-\tau^2/2}}{\tau\sqrt{2\pi}} \\ &\geq \frac{e^{c'log(p)}}{\sqrt{log(p)}} \text{ for some } c' > 0 \text{ since } \tau \ll \sqrt{log(p)} \\ &\to \infty.\end{aligned} \tag{B.32}$$

Now suppose $\frac{1}{16\theta^{'}(0)^2} < t < \rho^*_{\text{binary}}(\alpha)$. If $\epsilon = -\delta < 0$, then $\eta < \frac{\sqrt{2log(p)}\delta}{4\theta^{'}(0)} - \frac{(log(r)+x)}{\sqrt{r}}$. If $\eta \in (-2, \frac{\sqrt{2log(p)}\delta}{4\theta^{'}(0)} - \frac{(log(r)+x)}{\sqrt{r}})$ then since $\Phi(\eta) \geq \Phi(-2)$ we have by the same argument as in (B.31) that $e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 \to \infty$. Now suppose $\eta \leq -1$. Then once again using

the fact that $\Phi(\eta) = \overline{\Phi}(-\eta)$ and Lemma A.1 we have that

$$
\begin{aligned}
& e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 \\
& = e^{\{\frac{1}{2}(4\theta^{'}(0)+\epsilon)^2(A^*)^2-(4\theta^{'}(0)+\epsilon)(log(r)+x)A-4\theta^{'}(0)^2A^2r+ro(A^2)\}}\overline{\Phi}(-\eta) \\
& \geq (1-\frac{1}{\eta^2})\frac{\phi(\eta)}{-\eta}e^{\{\frac{1}{2}(4\theta^{'}(0)+\epsilon)^2(A^*)^2-(4\theta^{'}(0)+\epsilon)(log(r)+x)A-4\theta^{'}(0)^2A^2r+ro(A^2)\}} \\
& = \frac{(1-\frac{1}{\eta^2})}{-\eta}e^{\{log(p)(1-2(1-2\theta^{'}(0)\sqrt{t})^2)+\kappa'\}}
\end{aligned}
\tag{B.33}
$$

where $|\kappa'| \ll log(p)$. Now

$$
\begin{aligned}
\inf_{\frac{1}{16\theta^{'}(0)^2}<t<\rho^*_{\text{binary}}(\alpha)}\{1-2(1-2\theta^{'}(0)\sqrt{t})^2\} &\geq \inf_{t<\rho^*_{\text{binary}}(\alpha)}\{1-2(1-2\theta^{'}(0)\sqrt{t})^2\} \\
&= 1-2(1-2\theta^{'}(0)\sqrt{\rho^*_{\text{binary}}(\alpha)})^2 \\
&= 1-2(1-2\theta^{'}(0)\sqrt{\frac{(1-\sqrt{1-\alpha})^2}{4\theta^{'}(0)^2}})^2 \\
&= 1-2(1-(1-\sqrt{1-\alpha}))^2 = 2\alpha-1>0
\end{aligned}
$$

since $\alpha > \frac{1}{2}$. Hence from (B.33) we have that

$$
\begin{aligned}
e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 &\geq \frac{3}{-4\eta}e^{c''log(p)} \text{ for some } c''>0 \\
&\to \infty
\end{aligned}
$$

since $|\eta| \lesssim log(p)$. This completes the proof of $e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_1 \to \infty$.

Next we prove $e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_2 = O(1)$. To this end note, that by the Cauchy-Schwarz Inequality,

$$
\begin{aligned}
& S_2 \\
& \leq e^{-(4\theta^{'}(0)-\epsilon)(log(r)+x)A}(\mathbb{E}_0[e^{(8\theta^{'}(0)-2\epsilon)A^*V}]\mathbb{P}_0(|W_r-B_r|>(log(r)+x)))^{1/2} \text{ where } V\sim \mathrm{N}(0,1) \\
& \leq (e^{-(4\theta^{'}(0)-\epsilon)\frac{(log(r)+x)}{\sqrt{r}}\sqrt{2tlog(p)}+(8\theta^{'}(0)-2\epsilon)^2tlog(p)-\lambda x})^{1/2} \text{ by Equation (B.13)}
\end{aligned}
\tag{B.34}
$$

Hence from (B.34) and (B.29) we have that $e^{(4\theta^{'}(0)+\epsilon)Ar}(2\theta(-A))^{2r}S_2 \to 0$ since $x = a_{r,p}log(p)$ where $a_{r,p}$ was chosen to diverge at a slow enough rate. This completes the verification of (B.21) and hence proves the theorem. □

Proof of Theorem 2.9. We will provide proof for the lower bound in problem 2.4 where $\theta \in BC^2(0)$. Using Remark 2.1, the proof also holds for problem 2.13. To analyze the power of the Higher Criticism test, we need to define the following quantities. Let $\frac{1}{2} + \delta = \theta(A)$. Also define S_1 to be a generic $\mathrm{Bin}(r, \frac{1}{2} + \delta)$ random variable and let $\mathbb{B}_1, \overline{\mathbb{B}}_1$ respectively denote the distribution function and survival function of S_1. Then

$$\mathbb{B}_1(t) = \mathbb{P}(\frac{|S_1 - \frac{r}{2}|}{\sqrt{\frac{r}{4}}} \leq t), \overline{\mathbb{B}}_1(t) = 1 - \mathbb{B}_1(t) \ .$$

The proof of the rest of the theorem relies on the following lemma.

Lemma B.5. *Let $r \gg log(p)$ and $t > \rho^*_{\mathrm{logistic}}(\alpha)$. Then there exists $s \in [1, \sqrt{3log(p)}]$ such that*

1. $\frac{k}{\sqrt{p}} \frac{\overline{\mathbb{B}}_1(s) - \overline{\mathbb{B}}(s)}{\sqrt{\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s))}} \gg log(p)$
2. $\frac{(p-k)\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s)) + k\overline{\mathbb{B}}_1(s)(1-\overline{\mathbb{B}}_1(s))}{k^2(\overline{\mathbb{B}}_1(s) - \overline{\mathbb{B}}(s))^2} \to 0$

Now we return to the proof of the main result. For any $z \in [1, \sqrt{3log(p)}] \cap \mathbb{N}$, $\mathsf{T}_{\mathrm{HC}} \geq W_p(z)$ where $W_p(z) = \sqrt{p}\frac{\overline{\mathbb{F}}_p(z) - \overline{\mathbb{B}}(z)}{\sqrt{\overline{\mathbb{B}}(z)(1-\overline{\mathbb{B}}(z))}}$. Hence by Chebysev's inequality it suffices to prove that there exists $s \in [1, \sqrt{3log(p)}]$ such that uniformly in $\beta \in \Theta^A_k$, $\frac{\mathbb{E}_\beta(W_p(s))}{\sqrt{2loglog(p)}} \to \infty$ and $\frac{\mathrm{Var}_{\boldsymbol{\beta}}(W_p(s))}{(\mathbb{E}_{\boldsymbol{\beta}}(W_p(s)))^2} \to 0$ when $t > \rho^*_{\mathrm{logistic}}(\alpha)$. Fix $\boldsymbol{\beta}^* \in \Theta^A_k$; thus $\boldsymbol{\beta}^*$ has 0 in $p-k$ locations, A in k_1 locations (say) and $-A$ in $k - k_1 = k_2$ locations. Now note that by symmetry $\mathbb{P}(|\mathrm{Bin}(r, \frac{1}{2} + \delta) - \frac{r}{2}| > t) = \mathbb{P}(|\mathrm{Bin}(r, \frac{1}{2} - \delta) - \frac{r}{2}| > t)$ for all $t > 0$. Hence it is easy to show that irrespective of k_1, k_2, $\mathbb{E}_{\boldsymbol{\beta}^*}(W_p(s)) = \frac{k}{\sqrt{p}} \frac{\overline{\mathbb{B}}_1(s) - \overline{\mathbb{B}}(s)}{\sqrt{\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s))}}$ and $\mathrm{Var}_{\boldsymbol{\beta}^*}(W_p(s)) = \frac{p-k}{p} + \frac{k}{p}\frac{\overline{\mathbb{B}}_1(s)(1-\overline{\mathbb{B}}_1(s))}{\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s))}$. Hence to show $\frac{\mathbb{E}_\beta(W_p(s))}{log(p)} \to \infty$ it suffices to show $\frac{k}{\sqrt{p}} \frac{\overline{\mathbb{B}}_1(s) - \overline{\mathbb{B}}(s)}{\sqrt{\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s))}} \gg \sqrt{log(p)}$ which is true by item 1 of Lemma B.5. Similarly to show that $\frac{\mathrm{Var}_{\boldsymbol{\beta}}(W_p(s))}{(\mathbb{E}_{\boldsymbol{\beta}}(W_p(s)))^2} \to 0$ it suffices to show that $\frac{(p-k)\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s)) + k\overline{\mathbb{B}}_1(s)(1-\overline{\mathbb{B}}_1(s))}{k^2(\overline{\mathbb{B}}_1(s) - \overline{\mathbb{B}}(s))^2} \to 0$ which is also true by item 2 of Lemma B.5. This completes the proof. □

Proof of Lemma B.5. By inspecting the expressions, it suffices to prove $\frac{k}{\sqrt{p}} \frac{\overline{\mathbb{B}}_1(s) - \overline{\mathbb{B}}(s)}{\sqrt{\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s))}} \to \infty$ as some positive power of p. Put $s = \lfloor 2\sqrt{2qlog(p)} \rfloor$ where $q = \min\{4t\theta^{'}(0)^2, \frac{1}{4}\}$. By the

choice of q, $s \in [1, \sqrt{3log(p)}] \cap \mathbb{N}$. Now by the Berry-Esseen approximation and Mill's Ratio,

$$\begin{aligned}
\frac{k}{\sqrt{p}} \frac{\overline{\mathbb{B}}_1(s) - \overline{\mathbb{B}}(s)}{\sqrt{\overline{\mathbb{B}}(s)(1-\overline{\mathbb{B}}(s))}} &\approx p^{\frac{1}{2}-\alpha} \frac{\overline{\Phi}\left(s - \frac{r\delta}{\sqrt{\frac{r}{4}}}\right)}{\sqrt{\overline{\Phi}(s)}} \approx p^{\frac{1}{2}-\alpha} \frac{\overline{\Phi}\left(s - \frac{r\theta^{'}(0)A}{\sqrt{\frac{r}{4}}}\right)}{\sqrt{\overline{\Phi}(s)}} \\
&= p^{\frac{1}{2}-\alpha} \frac{\overline{\Phi}\left(\frac{\sqrt{2qlog(p)}\sqrt{r} - \sqrt{2tlog(p)}\theta^{'}(0)\sqrt{r}}{\sqrt{\frac{r}{4}}}\right)}{\sqrt{\overline{\Phi}\left(\frac{\sqrt{2qlog(p)}\sqrt{r}}{\sqrt{\frac{r}{4}}}\right)}} \\
&= p^{\frac{1}{2}-\alpha} \frac{\overline{\Phi}\left(\sqrt{8qlog(p)} - \sqrt{8tlog(p)}\theta^{'}(0)\right)}{\sqrt{\overline{\Phi}\left(\frac{\sqrt{2qlog(p)}}{\sqrt{\frac{1}{4}}}\right)}} \\
&\approx e^{\frac{1}{2}-\alpha-\frac{8}{2}log(p)(\sqrt{q}-\sqrt{t}\theta^{'}(0))^2+\frac{8}{4}qlog(p)} \\
&= p^{\frac{1}{2}-\alpha+2q-4(\sqrt{q}-\sqrt{t}\theta^{'}(0))^2} .
\end{aligned}$$

The exponent of p above is given by

$$\frac{1}{2} - \alpha + 2q - 4(\sqrt{q} - \sqrt{t}\theta^{'}(0))^2 =: f(q) \text{ say} .$$

The function $f(q)$ is maximized at $q = 4t\theta^{'}(0)^2$ for $t \leq \frac{1}{16\theta^{'}(0)^2}$. The maximum value is $(\frac{1}{2} - \alpha) + 4t\theta^{'}(0)^2 > 0$ since $t > \rho^*_{\text{binary}}(\alpha)$. For $t > \frac{1}{16\theta^{'}(0)^2}$ if we put $q = \frac{1}{4}$, then $f(q) = (1-\alpha) - (1 - 2\sqrt{t}\theta^{'}(0))^2 > 0$ since $t > \max\{\rho^*_{\text{binary}}(\alpha), \frac{1}{16\theta^{'}(0)^2}\}$. Hence taking $s = \sqrt{2qlog(p)}$ where $q = \min\{4t\theta^{'}(0)^2, \frac{1}{4}\}$ proves the lemma. □

Proof of Proposition 2.1. Set $V_{(j)} = |Z - \frac{r}{2}|_{(j)}$ so that

$$\begin{aligned}
&\sup_{t \in [V_{(p-j)}, V_{(p-j+1)})} \sqrt{p} \frac{\overline{\mathbb{F}}_p(t) - \overline{\mathbb{B}}(t)}{\sqrt{\overline{\mathbb{B}}(t)(1-\overline{\mathbb{B}}(t))}} \\
&\qquad = \sup_{t \in [V_{(p-j)}, V_{(p-j+1)})} \sqrt{p} \frac{\frac{j}{p} - \overline{\mathbb{B}}(t)}{\sqrt{\overline{\mathbb{B}}(t)(1-\overline{\mathbb{B}}(t))}} \\
&\qquad = \sqrt{p} \frac{\frac{j}{p} - \inf_{t \in [V_{(p-j)}, V_{(p-j+1)})} \overline{\mathbb{B}}(t)}{\sqrt{\inf_{t \in [V_{(p-j)}, V_{(p-j+1)})} \overline{\mathbb{B}}(t)(1 - \inf_{t \in [V_{(p-j)}, V_{(p-j+1)})} \overline{\mathbb{B}}(t))}} .
\end{aligned}$$

Now $\overline{\mathbb{B}}(t)$ is a decreasing function of t and thus $\inf_{t\in[V_{(p-j)},V_{(p-j+1)})}\overline{\mathbb{B}}(t)) \geq \overline{\mathbb{B}}(V_{(p-j+1)})) = q_{(j)}$. Therefore we obtain that

$$\begin{aligned}
&\sup_{t\in[V_{(p-j)},V_{(p-j+1)})}\sqrt{p}\frac{\overline{\mathbb{F}}_p(t)-\overline{\mathbb{B}}(t)}{\sqrt{\overline{\mathbb{B}}(t)(1-\overline{\mathbb{B}}(t))}} \\
&\qquad = \sqrt{p}\frac{\frac{j}{p}-\inf_{t\in[V_{(p-j)},V_{(p-j+1)})}\overline{\mathbb{B}}(t)}{\sqrt{\inf_{t\in[V_{(p-j)},V_{(p-j+1)})}\overline{\mathbb{B}}(t)(1-\inf_{t\in[V_{(p-j)},V_{(p-j+1)})}\overline{\mathbb{B}}(t))}} \\
&\qquad \leq \sqrt{p}\frac{\frac{j}{p}-q_{(j)}}{\sqrt{q_{(j)}(1-q_{(j)})}}
\end{aligned}$$

since $\frac{c-x}{\sqrt{x(1-x)}}$ is a decreasing function of $x \in [0,1]$ for $c \in [0,1]$ and the proof is done. □

Proof of Theorem 2.10. We will provide proof for the lower bound in problem 2.4 where $\theta \in BC^2(0)$. Using Remark 2.1, the proof also holds for problem 2.13. As in proof of Theorem 2.8, we denote by $B_{jr}^{'}$ the version of Brownian Motion approximating $W_{jr}^{'}$ where $W_{jr}^{'} = W_{jr}\sqrt{r}$ and we can choose $B_{jr}^{'}$ independent for $j = 1,\ldots,p$. Let $B_{jr} = \frac{B_{jr}^{'}}{\sqrt{r}}$. For any $t_p > 0$,

$$\begin{aligned}
\mathbb{P}(\max_{1\leq j\leq p}|W_{jr}| \leq t_p) &= \mathbb{P}(\max_{1\leq j\leq p}|W_{jr}-B_{jr}+B_{jr}| \leq t_p) \\
&\geq \mathbb{P}(\max_{1\leq j\leq p}|W_{jr}-B_{jr}| + \max_{1\leq j\leq p}|B_{jr}| \leq t_p) \\
&\geq \mathbb{P}\Big(\max_{1\leq j\leq p}|B_{jr}| \leq t_p - \frac{log(r)+x}{\sqrt{r}}\Big) + o(1)
\end{aligned}$$

for some $x > 0$. By a similar token we can show that $\mathbb{P}(\max_{1\leq j\leq p}|W_{jr}| \leq t_p) \leq \mathbb{P}(\max_{1\leq j\leq p}|B_{jr}| \leq t_p + \frac{log(r)+x}{\sqrt{r}}) + o(1)$ for the same x above. Now by Lemma 11 of Arias-Castro et al. (2011) we have that

$$\mathbb{P}(\max_{1\leq j\leq p}|B_{jr}| \leq \kappa_p + \frac{s}{\sqrt{2log(p)}}) \to e^{-e^{-s}}$$

as $p \to \infty$ where $\kappa_p = \sqrt{2log(p)} - \frac{loglog(p)+4\pi-4}{2\sqrt{2log(p)}}$. Hence if $r \gg (log(r))^2 log(p)$ then $\frac{log(r)+x}{\sqrt{r}} = \frac{o(1)}{\sqrt{2log(p)}}$ for appropriately chosen x. Therefore, by following the arguments of Lemma 11, Lemma 12 and proof of Theorem 5 of Arias-Castro et al. (2011) we have the result when $r \gg (log(r))^2 log(p)$ if we can choose x appropriately. We choose it to be the same as our choice in the proof of Theorem 2.8. To be precise, since $r \gg log(p)$, there exists a sequence

$a_{r,p} \to \infty$ such that $r \gg a_{r,p} log(p)$. Take $x = a_{r,p}$. We skip the rest of the details. □

Proof of Theorem 2.11. We divide the proof into proofs of lower bound and upper bound respectively.

Part 1 : Proof of Lower Bound For the purpose of brevity assume that $\boldsymbol{X} = [\boldsymbol{X}_1^t : \boldsymbol{X}_2^t]^t$ where $\boldsymbol{X}_1$ is an $n^* \times p$ matrix whose rows comprises exactly of the rows in Ω^* and $\boldsymbol{X}_2$ is an $n_* \times p$ matrix whose rows consists of the rows of $\boldsymbol{X}$ with more than one non-zero element in its support. Note that, this can always be achieved by a permutation of the rows of $\boldsymbol{X}$ and hence this does change the validity of the theorem. Let

$$f(\boldsymbol{X}_1, \boldsymbol{\beta}, \boldsymbol{\beta}') = \prod_{i \in \Omega^*} \left[\theta(\boldsymbol{x}_i^t \boldsymbol{\beta}) \theta(\boldsymbol{x}_i^t \boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t \boldsymbol{\beta}) \theta(-\boldsymbol{x}_i^t \boldsymbol{\beta}') \right]$$

and

$$f(\boldsymbol{X}_2, \boldsymbol{\beta}, \boldsymbol{\beta}') = \prod_{i \notin \Omega^*} \left[\theta(\boldsymbol{x}_i^t \boldsymbol{\beta}) \theta(\boldsymbol{x}_i^t \boldsymbol{\beta}') + \theta(-\boldsymbol{x}_i^t \boldsymbol{\beta}) \theta(-\boldsymbol{x}_i^t \boldsymbol{\beta}') \right].$$

Note that by Lemma B.4, we have that $f(\boldsymbol{X}_1, \boldsymbol{\beta}, \boldsymbol{\beta}') \leq [\theta^2(QA) + \theta^2(-QA)]^{n_*}$ for any realizations $\boldsymbol{\beta}, \boldsymbol{\beta}'$ from π:

$$\begin{aligned} \mathbb{E}_0(L_\pi^2) &= 2^{n^* + n_*} \iint f(\boldsymbol{X}_1, \boldsymbol{\beta}, \boldsymbol{\beta}') f(\boldsymbol{X}_2, \boldsymbol{\beta}, \boldsymbol{\beta}') d\pi(\boldsymbol{\beta}) d\pi(\boldsymbol{\beta}') \\ &\leq 2^{n_*} \left[\theta^2(QA) + \theta^2(-QA) \right]^{n_*} 2^{n^*} \iint f(\boldsymbol{X}_1, \boldsymbol{\beta}, \boldsymbol{\beta}') d\pi(\boldsymbol{\beta}) d\pi(\boldsymbol{\beta}'). \end{aligned} \tag{B.35}$$

Now, using $\theta(A) = \frac{1}{2} + \Delta$ we have that $A^2 \ll \frac{\sqrt{p}}{kr^*}$ implies $\Delta^2 \ll \frac{\sqrt{p}}{kr^*}$ since $\theta \in BC^2(0)$. Following the exact arguments as in the proof of lower bound Theorem 2.6, one has

$$\begin{aligned} 2^{n^*} \iint f(\boldsymbol{X}_1, \boldsymbol{\beta}, \boldsymbol{\beta}') d\pi(\boldsymbol{\beta}) d\pi(\boldsymbol{\beta}') &= \mathbb{E}_0 \left[\left(\frac{1 + 4\Delta^2}{1 - 4\Delta^2} \right)^{\sum_{j \in m_3} r_j} (1 - 4\Delta^2)^{\sum_{j \in m_1 \cap m_2} r_j} \right] \\ &= \mathbb{E}_0 \left[\prod_{\in m_1 \cap m_2} \left(\frac{1 + 4\Delta^2}{1 - 4\Delta^2} \right)^{r_j \mathbf{I}(j \in m_3)} (1 - 4\Delta^2)^{r_j} \right] \\ &= \mathbb{E}_0 \left[\prod_{j \in m_1 \cap m_2} \frac{1}{2} \left((1 + 4\Delta^2)^{r_j} + (1 - 4\Delta^2)^{r_j} \right) \right] \\ &\leq \mathbb{E}_0 \left[\left(\frac{1}{2} \right)^{|m_1 \cap m_2|} \left((1 + 4\Delta^2)^{r^*} + (1 - 4\Delta^2)^{r^*} \right)^{|m_1 \cap m_2|} \right] \end{aligned}$$

where the second to the last line follows since given $j \in m_1 \cap m_2$, $\mathbf{I}(j \in m_3) \sim \text{Bernoulli}(\frac{1}{2})$, independent for all j and the last line follows from noting that for any $\lambda \in (0,1)$, $(1+\lambda)^x + (1-\lambda)^x$ is an increasing function of $x \geq 1$. Hence following the same argument as in Theorem 2.6 after equation B.4, we have that there exists a constant $C > 0$ such that

$$2^{n^*} \iint f(\boldsymbol{X}_1, \boldsymbol{\beta}, \boldsymbol{\beta}^{'}) d\pi(\boldsymbol{\beta}) d\pi(\boldsymbol{\beta}^{'}) \leq \left(1 + C \frac{\frac{k^2 r^{*2} \Delta^4}{p}}{k}\right)^k \to 0 \tag{B.36}$$

since $\Delta \ll \sqrt{\frac{\sqrt{p}}{kr^*}}$. Hence, by B.35 and B.36 we have $\mathbb{E}_0(L_\pi^2) = 1 + o(1)$ if $2^{n_*} \left[\theta^2(QA) + \theta^2(-QA)\right]^{n_*} = 1 + o(1)$.

However, since $\theta \in BC^2(0)$, there exists a constant $C_1 > 0$ such that $\theta^2(QA) + \theta^2(-QA) \leq \frac{1}{2}(1 + C_1 Q^2 A^2)$. Hence

$$2^{n_*} \left[\theta^2(QA) + \theta^2(-QA)\right]^{n_*} = O\left(\left(1 + C_1 \frac{Q^2 A^2 n_*}{n_*}\right)^{n_*}\right)$$

Now by assumption, $A^2 \ll \frac{\sqrt{p}}{kr^*}$ and $\frac{Q^2 n_*}{r^*} \ll p^{\frac{1}{2}-\alpha} = \frac{k}{\sqrt{p}}$, one has that $C_1 Q^2 A^2 n_* \to 0$ and hence $2^{n_*} \left[\theta^2(QA) + \theta^2(-QA)\right]^{n_*} = 1 + o(1)$ as required. This completes the proof of the lower bound. □

Part 2 : Proof of Upper Bound We begin by noting that when $n_* = 0$, then the proof follows along the same lines as the power analysis argument of GLRT in Theorem 2.6 by using the fact $r_j \geq r_* \gg log(p)$ for all $j = 1, \ldots, p$. The proof then immediately follows by noting that the definition of the GLRT does not depend on n_* and solely depends on the observations corresponding to indices in Ω^*, *i.e*, on $(y_i, \boldsymbol{x}_i^t)_{i \in \Omega^*}$ and does not even consider the data corresponding to $\boldsymbol{X}_2$.

Proof of Theorem 2.12. The proof follows from Theorem 2.2 and is omitted. □

Proof of Theorem 2.13. We divide the proof into the proof of lower bound and upper bound respectively.

Part 1 : Proof of Lower Bound Define the intervals for $j = 1, \ldots, p$:

$$H_{p,j} = \Big(\frac{r_j}{2} - \sqrt{2log(p)}\sqrt{\frac{r_j}{4}}, \frac{r_j}{2} + \sqrt{2log(p)}\sqrt{\frac{r_j}{4}} \Big) \ . \tag{B.37}$$

and put

$$D = \{ Z_j \in H_{p,j},\ j = 1, \ldots, p \},\ Z_j = \sum_{i \in \Omega_j} y_i,\ l = 1, \ldots, p. \tag{B.38}$$

By Hölder's inequality it can be shown that for proving a lower bound it suffices to prove,

$$\mathbb{E}_0(L_\pi \, \mathbf{I}_{D^c}) = o(1), \qquad \mathbb{E}_0(L_\pi^2 \, \mathbf{I}_D) = 1 + o(1) \ . \tag{B.39}$$

We first prove the first equality of (B.39). Since $\{y_i, i \notin \Omega^*\}$ is independent of $\{Z_j, j = 1, \ldots, p\}$, we have by a calculation similar to proof of Theorem 2.8,

$$\begin{aligned}
\mathbb{E}_0(L_\pi \, \mathbf{I}_{D^c}) \leq & \binom{p}{k}^{-1} 2^{-k} \sum_{m_1, \xi_1} \Big[\sum_{j \in m_1^1} \mathbb{E}_0 \left(2^{r_j} \{ \frac{\theta(A)}{\theta(-A)} \}^{Z_j} \{ \theta(-A) \}^r \mathbf{I}(Z_j \in H_{p,j}^c) \right) \\
& + \sum_{j \in m_1^{-1}} \mathbb{E}_0 \left(2^{r_j} \{ \frac{\theta(A)}{\theta(-A)} \}^{r - Z_j} \{ \theta(-A) \}^r \mathbf{I}(Z_j \in H_{p,j}^c) \right) \\
& + \sum_{j \in m_1^c} 2^{r_j} (\frac{1}{2})^{r_j} \mathbb{P}(Z_j \in H_{p,j}^c) \Big] \\
= & \binom{p}{k}^{-1} \sum_{m_1} \Big[\sum_{j \in m_1} (2\theta(-A))^{r_j} \mathbb{E}_0 \Big(\left(\frac{\theta(A)}{\theta(-A)} \right)^{Z_j} \mathbf{I}(Z_j \in H_{p,j}^c) \Big) + \sum_{j \in m_1^c} \mathbb{P}(Z_j \in H_{p,j}^c) \Big].
\end{aligned} \tag{B.40}$$

Now note that, by the same argument as proof of B.11, we have by an application of Lemma B.2,

$$\sum_{j \in m_1^c} \mathbb{P}(Z_j \in H_{p,j}^c) \leq 2 \sum_{j \in m_1^c} \frac{e^{-log(p)}}{\epsilon_j \sqrt{r_j}} e^{r_j \epsilon_j^2 - r_j \epsilon_j}$$

where $\epsilon_j = \frac{2\sqrt{\frac{r_j}{4}}\sqrt{2log(p)} - 1}{r_j - 1}$. Hence there exists a constant $C > 0$ which does not depend on j such that $\epsilon_j \leq C\sqrt{\frac{2log(p)}{r_j}}$. Therefore, $r_j \epsilon_j^2 - r_j \epsilon_j \leq C\sqrt{2log(p)}(C\sqrt{2log(p)} - r_j) \leq C\sqrt{2log(p)}(C\sqrt{2log(p)} - r_*)$. Also,there exists a constant $c > 0$, not depending on j such

that $\epsilon_j\sqrt{r_j} \geq c\sqrt{2log(p)}$. Thus

$$\sum_{j\in m_1^c} \mathbb{P}(Z_j \in H_{p,j}^c) \leq 2\frac{C\sqrt{2log(p)}(C\sqrt{2log(p)} - r_*)(p-k)}{pc\sqrt{2log(p)}} \to 0 \tag{B.41}$$

since $r_* \gg log(p)$.

Next we control the term $\sum_{j\in m_1}(2\theta(-A))^{r_j}\mathbb{E}_0\Big(\Big(\frac{\theta(A)}{\theta(-A)}\Big)^{Z_j}\mathbf{I}(Z_j \in H_{p,j}^c)\Big)$. To this end note that, by a proof similar to that of controlling B.12, one has using $r_* \gg log(p)$ and $t < \rho^*_{\text{binary}}(\alpha)$ that there exists a sequence of real numbers $\lambda_p = o(1)$ which does not depend j such that $k(2\theta(-A))^{r_j}\mathbb{E}_0\Big(\Big(\frac{\theta(A)}{\theta(-A)}\Big)^{Z_j}\mathbf{I}(Z_j \in H_{p,j}^c)\Big) \leq \lambda_p$. In particular, this sequence can be taken to be polynomially p, as the proof of Theorem 2.8 suggests. This implies that

$$\begin{aligned}
&\sum_{j\in m_1}(2\theta(-A))^{r_j}\mathbb{E}_0\Big(\Big(\frac{\theta(A)}{\theta(-A)}\Big)^{Z_j}\mathbf{I}(Z_j \in H_{p,j}^c)\Big) \\
&= \frac{1}{k}\sum_{j\in m_1} k(2\theta(-A))^{r_j}\mathbb{E}_0\Big(\Big(\frac{\theta(A)}{\theta(-A)}\Big)^{Z_j}\mathbf{I}(Z_j \in H_{p,j}^c)\Big) \leq \lambda_p.
\end{aligned} \tag{B.42}$$

Hence, by B.40, we have using B.41 and B.42 that

$$\mathbb{E}_0(L_\pi \mathbf{I}_{D^c}) = o(1)$$

as required. This completes the proof of the first equality of (B.39). Next we prove the second claim of (B.39). Arguing similarly as in analysis of equation B.20 in proof of Theorem 2.8 and using Lemma B.4 we have that

$$\begin{aligned}
&\mathbb{E}_0(L_\pi^2 \mathbf{I}_D) \\
&\leq 2^{n_*}\left[\theta^2(QA) + \theta^2(-QA)\right]^{n_*} \\
&\times \mathbb{E}_{m_1\cap m_2}\Big\{\prod_{j=1}^{p}\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^{r_j}\mathbb{P}(Z_j \in H_{p,j}) \\
&\quad + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_j}\mathbf{I}(Z_j \in H_{p,j})\Big](2\theta(-A))^{2r_j}\Big)\Big]^{\mathbf{I}(j\in m_1\cap m_2)}\Big\}.
\end{aligned}$$

As in proof of claim B.21 in Theorem 2.8, we have that for any $j = 1, \ldots, p$,

$$\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^{r_j}\mathbb{P}(Z_j \in H_{p,j}) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_j}\mathbf{I}(Z_j \in H_{p,j})\Big](2\theta(-A))^{2r_j}\Big)\Big] \geq 1$$

since $r_* \gg log(p)$ and $t < \rho^*_{\text{binary}}(\alpha)$. Let

$$j^* = \underset{j}{\text{argmax}}\{\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^{r_j}\mathbb{P}(Z_j \in H_{p,j}) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z_j}\mathbf{I}(Z_j \in H_{p,j})\Big](2\theta(-A))^{2r_j}\Big)\}$$

and let $r = r_{j^*},\ Z = Z_{j^*}, H_p = H_{p,j^*}$. Hence, one has with $U \sim Bin(k, \frac{k}{p-k})$ and $\varphi_{n,p} = 2^{n_*}\left[\theta^2(QA) + \theta^2(-QA)\right]^{n_*}$, that

$$\begin{aligned}
&\mathbb{E}_0(L_\pi^2\mathbf{I}_D)\\
&\le \varphi_{n,p}\mathbb{E}_{m_1\cap m_2}\Big\{\prod_{j=1}^{p}\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^{r}\mathbb{P}(Z \in H_p)\\
&\qquad + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z}\mathbf{I}(Z \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big]^{\mathbf{I}(j\in m_1\cap m_2)}\Big\}\\
&= \varphi_{n,p}\mathbb{E}_{m_1\cap m_2}\Big\{\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^{r}\mathbb{P}(Z \in H_p)\\
&\qquad + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z}\mathbf{I}(Z \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big]^{|m_1\cap m_2|}\Big\}\\
&\le \varphi_{n,p}\mathbb{E}_U\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^{r}\mathbb{P}(Z \in H_p) + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z}\mathbf{I}(Z \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big]^{U}\\
&= \varphi_{n,p}\Big\{1 + \frac{k}{p-k}\Big(\Big[\frac{1}{2}\Big(\{4\theta(A)\theta(-A)\}^{r}\mathbb{P}(Z \in H_p)\\
&\qquad + \mathbb{E}_0\Big[\Big\{\frac{\theta(A)}{\theta(-A)}\Big\}^{2Z}\mathbf{I}(Z \in H_p)\Big](2\theta(-A))^{2r}\Big)\Big] - 1\Big)\Big\}
\end{aligned} \tag{B.43}$$

where the second to the last line follows from Lemma B.1. Using the fact that $r = r_{j^*} \ge r_* \gg log(p)$, one has by similar argument as in the proof of B.22 in Theorem 2.8 that

$$\mathbb{E}_0(L_\pi^2\mathbf{I}_D) = \varphi_{n,p}(1 + o(1))$$

when $t < \rho^*_{\text{binary}}(\alpha)$. Hence the verification of the second claim in (B.39) will be complete if we prove that $\varphi_{n,p} = 1 + o(1)$. Now, since $\theta \in BC^2(0)$, there exists a constant $C_1 > 0$ such that $\theta^2(QA) + \theta^2(-QA) \le \frac{1}{2}(1 + C_1Q^2A^2)$. Hence

$$\varphi_{n,p} = 2^{n_*}\left[\theta^2(QA) + \theta^2(-QA)\right]^{n_*} = O\left(\left(1 + C_1\frac{Q^2A^2n_*}{n_*}\right)^{n_*}\right)$$

Now by assumption, $A^2 = \frac{2tlog(p)}{r^*}$ with $t < \rho^*_{\text{binary}}(\alpha)$ and $\frac{Q^2n_*}{r^*} \ll log(p)$, one has that $C_1Q^2A^2n_* \to 0$ and hence $2^{n_*}\left[\theta^2(QA) + \theta^2(-QA)\right]^{n_*} = 1 + o(1)$. This justifies second

equality of (B.39) and hence completes the proof of Theorem 2.13.

Part 2 : Proof of Upper Bound We begin by noting that when $n_* = 0$, then the proof follows by very similar way as the power analysis argument of the Higher Criticism test in Theorem 2.9 by using the fact $r_j \geq r_* \gg log(p)$ for all $j = 1, \ldots, p$ and hence we omit the details. The proof then immediately follows by noting that the definition of the Higher Criticism test does not depend on n_* and solely depends on the observations corresponding to indices in Ω^*, *i.e*, on $(y_i, \boldsymbol{x}_i^t)_{i \in \Omega^*}$ and does not even consider the data corresponding to $\boldsymbol{X}_2$. □

Appendix C

Proofs for Chapter 3

Preliminary Results on Projection Kernels

In the following we review some results on projection operators and kernels based on Robins et al. (2013). Throughout μ, ν will stand for arbitrary σ-finite measures on $[0,1]^d$. For any measure ν on $[0,1]^d$, *kernel operator* $K : L_2(\nu) \to L_2(\nu)$ is defined to take the form $(Kf)(x) = \int \bar{K}(x,y)h(y)\, d\nu(y)$ for some measurable function $\bar{K} : ([0,1]^d)^2 \to \mathbb{R}$. By abuse of notation, let us denote the operator K and the *kernel* $\bar{K}$ with the same symbol: $K = \bar{K}$. By a "*weighted projection*" in $L_2(\mu)$ onto a closed subspace L with weight function w we will mean the map $\Pi : L_2(\mu) \to L$ given by

$$\Pi(g) = argmin_{l \in L} \int (g-l)^2\, w\, d\mu.$$

If the functions w of interest are bounded away from 0 and ∞, this map always exists. It can be shown that the projection map is also determined by the following two constraints: $\Pi(g) \in L$ and the orthogonality equation

$$\int (g - \Pi(g)) l\, w\, d\mu = 0, \qquad \forall l \in L.$$

A weighted projection is often said to have a "*kernel representation with kernel*" Π if for all $g \in L_2(\mu)$,

$$\Pi(g)(x_1) = \int \Pi(x_1, x_2) g(x_2) w(x_2)\, d\mu(x_2).$$

For example the projections in our case will have kernel representations with kernels $K_{f,k}$. We note that, a weighted projection is basically an orthogonal projection onto L in

the space $L_2(\nu)$ for the changed measure ν defined by $d\nu = w\,d\mu$. Also as an operator on $L_2(\nu)$ it has kernel Π. However, as a operator on $L_2(\mu)$ the weighted projection has kernel $(x_1, x_2) \mapsto \Pi(x_1, x_2)w(x_2)$. This ambiguity is unavoidable if one needs to work with multiple weight functions, both estimated and "true" ones. We note that the kernel of an orthogonal projection is symmetric in its arguments and hence with the preceding convention the "kernel of a weighted projection" is symmetric in its arguments as well. Finally, it is worth noting that not all projections have kernels. However, projections onto finite-dimensional subspaces can be represented as kernels as is shown by the next simple lemma.

Lemma C.1. *Let $e_1, \ldots, e_k$ be arbitrary linearly independent elements spanning a subspace L of $L_2(\mu)$. Then the weighted projection onto L relative to the weight function w has kernel of projection*

$$\Pi(x_1, x_2) = \sum_i \sum_j (C^{-1})_{ij} e_i(x_1) e_j(x_2),$$

for $C_{k\times k}$ with $C_{ij} = \int e_i e_j w\,d\mu$.

Proof. Since one can always perform a change of measure from μ to ν by $d\nu = w\,d\mu$, it suffices to prove the lemma for $w = 1$. Then if $\Pi(g) = \sum_i \gamma_i e_i$, the orthogonality relationship encoded by $g - \Pi(g) \perp e_j$ imply that $\sum_i \gamma_i C_{ij} = \int g e_j\,d\nu$ for $j = 1, \ldots, k$. Inverting this system the equations implies that $\gamma_i = \sum_j (C^{-1})_{ij} \int g e_j\,d\nu$ for all i. Finally, plugging this into $g = \sum_i \gamma_i e_i$ and exchanging the order of summation and integration completes the proof. □

Next we note that, although by definition, an orthogonal projection in $L_2(\mu)$ has operator norm 1, its square $L_2(\mu\times\mu)$-norm $\int\int \Pi^2\,d(\mu\times\mu)$ is equal to the dimension of its projection space. This simple result is collected in the next lemma.

Lemma C.2. *The kernel of an orthogonal projection on a k-dimensional subspace of $L_2(\mu)$ has square $L_2(\mu \times \mu)$-norm $\int\int \Pi^2\,d(\mu \times \mu) = k$.*

Proof. We can write the kernel using Lemma C.1 relative to an orthonormal basis $e_1, \ldots, e_k$

of the projection space. Since, in such cases, $C = I$, one has

$$\int\int \Pi^2 \, d(\mu \times \mu) = \sum_i \sum_j \int\int e_i(x_1)e_i(x_2)e_j(x_1)e_j(x_2) \, d\mu(x_1) \, d\mu(x_2).$$

The cross terms in the above sum vanish by orthogonality, while the diagonal elements are equal to 1 by orthonormality. □

The next part of the discussion will be required for calculation of variance of HOIFs. The square norm of a projection kernel can typically be written as $\int \Pi(x,x) \, d\mu(x)$. The projection property encoded by $\Pi^2 = \Pi$ of a kernel operator Π on $L_2(\mu)$ can be also be expressed as

$$\int \Pi(x_1,x_2)\Pi(x_2,x_3) \, d\mu(x_2) = \Pi(x_1,x_3), \qquad \text{a.e. } (x_1,x_3). \tag{C.1}$$

Suppose the above holds for every $x_1 = x_3$. Then by integration one has that $\int\int \Pi^2 \, d(\mu \times \mu) = \int \Pi(x,x) \, d\mu(x)$. We will throughout assume that (C.1) is valid for every x_1, x_3. In particular, this will be assumed to hold on $\{(x_1,x_3) : x_1 = x_3\}$, which is typically a $\mu-$null set. In particular, this can be shown to hold kernels in Lemma C.1. This is the content of the next lemma.

Lemma C.3. *If $\Pi_1, \ldots, \Pi_{m-1}$ are kernels of orthogonal projections in $L_2(\mu)$ that satisfy (C.1) identically, then*

$$\int \cdots \int \prod_{i=1}^{m-1} \Pi_i^2(x_i, x_{i+1}) \, d\mu(x_1) \cdots d\mu(x_m) \leq \|\mu\| \prod_{i=1}^{m-1} \sup_x \Pi_i(x,x)$$

where $\|\mu\|$ denotes the L_1 or total variation norm of μ.

Proof. First note that, by Equation (C.1) $\int \Pi_i(x,y)^2 \, d\mu(y) = \Pi_i(x,x)$ for every x. Applying this to the integral with respect to x_m of the multiple integral in the lemma turns the m-fold integral into an $(m-1)$fold integral of the function $\prod_{i=1}^{m-2} \Pi_i^2(x_i, x_{i+1})\Pi_{m-1}(x_{m-1}, x_{m-1})$. Now we can bound the factor $\Pi_{m-1}(x_{m-1}, x_{m-1})$ by its supremum over x_{m-1}. This leaves us with a $(m-1)$fold integral of exactly the same type as before times the supremum we obtained. Repeating this argument until the only remaining integral is $\int \Pi_1(x_1,x_1) \, d\mu(x_1)$, which is bounded above by $\sup_x \Pi_1(x,x)\|\mu\|$,

completes the proof of the lemma. □

By Lemma C.3 above one has that under (C.1), the square norms of the products of projection kernels are controlled by their values on the diagonal type of terms. Since one often needs to work with weighted projection kernels, the following lemma is important where one shows that these values do not differ significantly for weighted projections with different weights either.

Lemma C.4. *Let the weighted projections in $L_2(\mu)$ onto a finite-dimensional space L relative to the weight functions v and w have kernels Π_v and Π_w satisfying (C.1) identically. Then for every x one has,*

$$\Pi_v(x,x) \leq \left\| \frac{w}{v} \right\|_\infty \Pi_w(x,x).$$

Proof. Fix a basis $e_1, \ldots, e_k$ of L. Then by Lemma C.1 one can represent the kernels by $\Pi_v(x,y) = \vec{e}_k(x)^T C_v^{-1} \vec{e}_k(y)$ where C_v is the matrix with (i,j)th element given by $\int e_i e_j v \, d\mu$. A similar expression holds for for Π_w. Here $\vec{e}_k = (e_1, \ldots, e_k)^T$. Now let us choose $e_1, \ldots, e_k$ to be orthonormal in $L_2(w)$. Then the matrix C_w is the $k \times k$ identity matrix. Therefore, The quotient $\Pi_v(x,x)/\Pi_w(x,x)$ is $z^T C_v^{-1} z / z^T z$ for some $z \in \mathbb{R}^k$. Hence it suffices to upper bound this quotient uniformly in $z \in \mathbb{R}^k$. But $\sup_{z \in \mathbb{R}^k} z^T C_v^{-1} z / z^T z = \frac{1}{\lambda_{\min}(C_v)}$ where $\lambda_{\min}(C_v)$ is the minimum eigenvalue of C_v. However,

$$z^T C_v z = \int \left(\sum_{i=1}^k z_i e_k \right)^2 v \, d\mu \geq \inf_x \frac{v}{w}(x) \int \left(\sum_{i=1}^k z_i e_k \right)^2 w \, d\nu = \inf_x \frac{v}{w}(x) \, z^T z.$$

Therefore $\lambda_{\min}(C_v)$ is bounded below the minimum value of v/w which in turn completes the proof. □

Proof of Main Results

Proof of Proposition 3.2. Our proof of closely follows the proof of Proposition 3.1 from Li et al. (2011) and similar results in Robins et al. (2008). In particular it follows directly from proofs of Theorem 1 and Theorem 5 in Li et al. (2011) that

$$TB_k(\theta) = \mathbb{E}_\theta[(\Pi_\theta^\perp[(b(\boldsymbol{X}) - \hat{b}(\boldsymbol{X}))|\overline{\mathbf{Z}}_{\hat{f},k}]) \times (\Pi_\theta^\perp[(p(\boldsymbol{X}) - \hat{p}(\boldsymbol{X}))|\overline{\mathbf{Z}}_{\hat{f},k}])]$$

and

$$\begin{aligned} EB_{m,k}(\theta) = (-1)^m \mathbb{E}_\theta[(b(\boldsymbol{X}) - \hat{b}(\boldsymbol{X}))\overline{\mathbf{Z}}_{\hat{f},k}^T] \\ \times \{\mathbb{E}_\theta[\overline{\mathbf{Z}}_{\hat{f},k}\overline{\mathbf{Z}}_{\hat{f},k}^T]^{-1} - I_{k\times k}\} \\ \times \{\mathbb{E}_\theta[\overline{\mathbf{Z}}_{\hat{f},k}\overline{\mathbf{Z}}_{\hat{f},k}^T] - I_{k\times k}\}^{m-2} \\ \times \mathbb{E}_\theta[\overline{\mathbf{Z}}_{\hat{f},k}(p(\boldsymbol{X}) - \hat{p}(\boldsymbol{X}))] \end{aligned}$$

where all the expectations are to be understood conditional on the training sample. The fact that $EB_m = \sup_\theta(EB_{m,k})$ is of the claimed order follows directly along the lines of proof in Robins et al. (2008) provided now $\hat{b}, \hat{p}, \hat{f}$ attain optimal rates of convergence over Sobolev classes of respective smoothnesses. For the order of truncation bias we will use the approximation property of the trigonometric basis in Sobolev classes. In particular, by Cauchy-Schwartz inequality,

$$TB_k^2(\theta) \leq \mathbb{E}_\theta\left[(\Pi_\theta^\perp[(b(\boldsymbol{X}) - \hat{b}(\boldsymbol{X}))|\overline{\mathbf{Z}}_{\hat{f},k}])^2\right]\mathbb{E}_\theta\left[(\Pi_\theta^\perp[(p(\boldsymbol{X}) - \hat{p}(\boldsymbol{X}))|\overline{\mathbf{Z}}_{\hat{f},k}])^2\right].$$

However, for any $h \in \{b, p\}$, we have by optimal approximation property of trigonometric basis in Sobolev ellipsoids

$$\begin{aligned} &\mathbb{E}_\theta\left[(\Pi_\theta^\perp[(h(\boldsymbol{X}) - \hat{h}(\boldsymbol{X}))|\overline{\mathbf{Z}}_{\hat{f},k}])^2\right] \\ &= \mathbb{E}_\theta\left[(\Pi_\theta^\perp[(b(\boldsymbol{X}) - \hat{b}(\boldsymbol{X}))|\bar{\phi}_k(\boldsymbol{X})])^2\right] \\ &= \inf_{\zeta_1,\ldots,\zeta_k}\int(h(\boldsymbol{x}) - \hat{h}(\boldsymbol{x}) - \sum_{l=1}^k \zeta_l\phi_l(\boldsymbol{x}))^2 f(\boldsymbol{x})d\boldsymbol{x} \\ &\leq \|f\|_\infty \inf_{\zeta_1,\ldots,\zeta_k}\int(h(\boldsymbol{x}) - \hat{h}(\boldsymbol{x}) - \sum_{l=1}^k \zeta_l\phi_l(\boldsymbol{x}))^2 d\boldsymbol{x} \\ &= O_p(k^{-2\beta_h}) \end{aligned}$$

and thereby completing the proof for truncation bias. As for the variance, it can be shown by analysis similar to the variance calculations in Li et al. (2011); Robins et al. (2008) and Lemma C.3 that

$$\epsilon_{i_1}\overline{\mathbf{Z}}_{f,k}(\boldsymbol{X}_{i_1})^T\prod_{r=3}^j\left(\overline{\mathbf{Z}}_{f,k}(\boldsymbol{X}_{i_r})\overline{\mathbf{Z}}_{f,k}(\boldsymbol{X}_{i_r})^T - I_{k\times k}\right)\overline{\mathbf{Z}}_{f,k}(\boldsymbol{X}_{i_2})\Delta_{i_2}$$

has variance under θ bounded by $sup_x(K_{f,k}(x,x))^{m-1}$ when f is the Lebesgue density. However when f is the Lebesgue density, for odd $k \geq 3$, the trigonometric basis has $K_{f,k}(x,x) = \frac{k+1}{2}$ for all x and for even k, $K_{f,k}(x,x) \leq \frac{k}{4} + 1$ by simple sine cosine calculations. Therefore $sup_x(K_{f,k}(x,x))^{m-1} \leq k^{m-1}$. Combining this with Lemma C.4, the variance of $H^{(k)}_{j,j,i_j^-}$ is also of the order of k^{m-1} under general f, since under the assumptions of the proposition $\|\frac{f}{\hat{f}}\|_\infty$ is bounded away from 0 and ∞. □

Proof of Proposition 3.4. First note that there exists g^* such that

$$\begin{aligned} & \Pi\left[\mathbb{E}\mathbb{S}_1|\mathcal{V}^1_{\alpha\eta}\right] \\ = \; & V^1_{\alpha\eta}\left(g^*/n\left(n-1\right)\right) \\ = \; & \frac{1}{n}\sum_i \frac{1}{(n-1)}\epsilon_i \mathbf{Z}_{ki}^T g^*\left(\boldsymbol{X}_i\right)\Delta_i + \frac{1}{n\left(n-1\right)}\sum_{i\neq j}\epsilon_i \mathbf{Z}_{ki}^T \Delta_j g^*\left(\boldsymbol{X}_j\right) \end{aligned}$$

such that it satisfies for all $g \in L_2(F_{\boldsymbol{X}})$

$$n\mathbb{E}\left[\epsilon_i \Delta_i V^1_{\alpha\eta}\left(g\right)\right] = \mathbb{E}\left[V^1_{\alpha\eta}\left(g^*\right)V^1_{\alpha\eta}\left(g\right)\right].$$

Hence

$$\begin{aligned} n\mathbb{E}\left[\mathbf{Z}_{ki}^T g\left(\boldsymbol{X}_i\right)\right] &= \frac{1}{(n-1)}\mathbb{E}\left[g^*\left(\boldsymbol{X}_i\right)^T \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T g\left(\boldsymbol{X}_i\right)\right] + \mathbb{E}\left[g^*\left(\boldsymbol{X}_i\right)^T \mathbf{Z}_{kj}\mathbf{Z}_{kj}^T g\left(\boldsymbol{X}_i\right)\right] \\ &= \frac{1}{(n-1)}\mathbb{E}\left[g^*\left(\boldsymbol{X}_i\right)^T \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T g\left(\boldsymbol{X}_i\right)\right] + \mathbb{E}\left[g^*\left(\boldsymbol{X}_i\right)^T g\left(\boldsymbol{X}_i\right)\right]. \end{aligned}$$

Since this holds for all $g \in L_2(F_{\boldsymbol{X}})$ we have

$$\begin{aligned} n\mathbf{Z}_{ki}^T &= \frac{1}{(n-1)} g^*\left(\boldsymbol{X}_i\right)^T \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T + g^*\left(\boldsymbol{X}_i\right)^T \\ & \quad or \\ g^*\left(\boldsymbol{X}_i\right) &= \left\{I + \mathbf{Z}_{ki}\mathbf{Z}_{ki}^T \frac{1}{(n-1)}\right\}^{-1} n\mathbf{Z}_{ki} \\ &= \left\{I - \frac{\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T}{\{n-1\} + \mathbf{Z}_{ki}^T\mathbf{Z}_{ki}}\right\} n\mathbf{Z}_{ki}. \end{aligned}$$

Now

$$\frac{1}{(n-1)}\mathbf{Z}_{ki}^{T}g^{*}\left(\boldsymbol{X}_{i}\right)=\frac{n}{(n-1)}\left[\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}-\frac{\left\{\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}\right\}^{2}}{\left\{n-1\right\}+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right]$$
$$=\frac{\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}{n-1}\left\{\frac{n\left(\mathbf{n}-\mathbf{1}\right)}{\left\{n-1\right\}+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right\}=\left\{\frac{n\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}{\left\{n-1\right\}+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right\}.$$

Hence, $\mathbf{Z}_{ki}^{T}g^{*}\left(\boldsymbol{X}_{j}\right)=n\mathbf{Z}_{ki}{}^{T}\mathbf{Z}_{kj}\left\{1-\frac{\mathbf{Z}_{kj}{}^{T}\mathbf{Z}_{kj}}{\{n-1\}+\mathbf{Z}_{kj}{}^{T}\mathbf{Z}_{kj}}\right\}=\mathbf{Z}_{ki}{}^{T}\mathbf{Z}_{kj}\frac{n(\mathbf{n}-\mathbf{1})}{\{n-1\}+\mathbf{Z}_{kj}{}^{T}\mathbf{Z}_{kj}}$. Plugging the optimal g^{*} in we get

$$\begin{aligned}
& V_{\alpha\eta}^{1}\left(g^{*}/n\left(n-1\right)\right)\\
= & \frac{1}{n}\sum_{i}\frac{1}{(n-1)}\epsilon_{i}\mathbf{Z}_{ki}^{T}g^{*}\left(\boldsymbol{X}_{i}\right)\Delta_{i}+\frac{1}{n\left(n-1\right)}\sum_{i\neq j}\epsilon_{i}\mathbf{Z}_{ki}^{T}\Delta_{j}g^{*}\left(\boldsymbol{X}_{j}\right)\\
= & \frac{1}{n}\sum_{i}\epsilon_{i}\frac{n\mathbf{Z}_{ki}{}^{T}\mathbf{Z}_{ki}}{\left\{n-1\right\}+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\Delta_{i}+\frac{1}{n\left(n-1\right)}\sum_{i\neq j}\epsilon_{i}\mathbf{Z}_{ki}{}^{T}\mathbf{Z}_{kj}\frac{n\left(\mathbf{n}-\mathbf{1}\right)}{\left\{n-1\right\}+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\Delta_{j}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{ES}_{2} \equiv & \mathbb{ES}_{1}-\Pi\left[\mathbb{ES}_{1}|\mathcal{V}_{\alpha\eta}^{1}\right]\\
= & \frac{1}{n}\sum_{i}\epsilon_{i}\left\{n-\frac{n\mathbf{Z}_{ki}{}^{T}\mathbf{Z}_{ki}}{n-1+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right\}\Delta_{i}-\frac{1}{n\left(n-1\right)}\sum_{i\neq j}\epsilon_{i}\mathbf{Z}_{ki}{}^{T}\mathbf{Z}_{kj}\frac{n\left(n-1\right)}{n-1+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\Delta_{j}\\
= & \frac{1}{n}\sum_{i}\epsilon_{i}\left\{\frac{n\left(n-1\right)}{n-1+\mathbf{Z}_{ki}^{T}\mathbf{Z}_{ki}}\right\}\Delta_{i}-\frac{1}{n\left(n-1\right)}\sum_{i\neq j}\epsilon_{i}\mathbf{Z}_{ki}{}^{T}\mathbf{Z}_{kj}\frac{n\left(n-1\right)}{n-1+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\Delta_{j}.
\end{aligned}$$

□

Proof of Proposition 3.5. Let $\widetilde{V}^{1:2}_{\alpha\eta\omega}(\gamma) = n(n-1)(n-2)V^{1:2}_{\alpha\eta\omega}(\gamma)$. First note that,

$$
\begin{aligned}
& \mathbb{E}\left[V^{1:2}_{\alpha\eta\omega}(\gamma^*)\,\widetilde{V}^{1:2}_{\alpha\eta\omega}(r,a)\right] \\
= \quad & \frac{1}{(n-1)(n-2)}\mathbb{E}\left\{\begin{array}{c}\left(\frac{n-1}{(n-1)+\mathbf{Z}_{ki}^T\mathbf{Z}_{ki}}\right)^2 r(\boldsymbol{X}_j)^T \\ \mathbb{E}\left[\left(a(\boldsymbol{X})\,\mathbf{Z}_k\mathbf{Z}_k^T\right)\right]\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\mathbb{E}\left[\left(\mathbf{Z}_k\mathbf{Z}_k^T\gamma^*(\boldsymbol{X}_j,\boldsymbol{X})\right)\right]\end{array}\right\} \\
& -\frac{1}{(n-2)}\mathbb{E}\left[\begin{array}{c} r(\boldsymbol{X}_j)^T\mathbb{E}\left[\left(a(\boldsymbol{X})\left[\mathbf{Z}_k\mathbf{Z}_k^T - I\right]\right)\right] \\ \left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}\mathbb{E}\left[\left(\mathbf{Z}_k\mathbf{Z}_k^T\gamma^*(\boldsymbol{X}_j,\boldsymbol{X})\right)\right]\end{array}\right] \\
& -\frac{1}{(n-2)}\mathbb{E}\left[\begin{array}{c} r(\boldsymbol{X}_j)^T\mathbb{E}\left[a(\boldsymbol{X}_i)\,\mathbf{Z}_{ki}\mathbf{Z}_{ki}^T\right] \\ \left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}\mathbb{E}\left[\left(\mathbf{Z}_k\mathbf{Z}_k^T\gamma^*(\boldsymbol{X}_j,\boldsymbol{X})\right)\right]\end{array}\right] \\
& +\frac{1}{(n-2)}\mathbb{E}\left[r(\boldsymbol{X}_j)^T\mathbb{E}\left[\,a(\boldsymbol{X})\,\mathbf{Z}_k\mathbf{Z}_k^T\gamma^*(\boldsymbol{X}_j,\boldsymbol{X})\right]\right] \\
& +\frac{1}{(n-2)}\mathbb{E}\left[\begin{array}{c} r(\boldsymbol{X}_j)^T\mathbb{E}\left[\left(a(\boldsymbol{X})\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\right)\right] \\ \left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}^2\mathbb{E}\left[\left(\mathbf{Z}_k\mathbf{Z}_k^T\gamma^*(\boldsymbol{X}_j,\boldsymbol{X})\right)\right]\end{array}\right] \\
& +\frac{1}{(n-2)}\mathbb{E}\left[r^T(\boldsymbol{X}_s)\,\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T\gamma^*(\boldsymbol{X}_s,\boldsymbol{X})\,a(\boldsymbol{X})\right] + \mathbb{E}\left[r(\boldsymbol{X}_j)^T\gamma^*(\boldsymbol{X}_s,\boldsymbol{X})\,a(\boldsymbol{X})\right].
\end{aligned}
$$

Now, by definition of a projection, one has for all r, a with finite variance and a having mean 0,

$$\mathbb{E}[\widetilde{V}^{1:2}_{\alpha\eta\omega}(r,a)\,U_{33}] = \mathbb{E}[\widetilde{V}^{1:2}_{\alpha\eta\omega}(r,a)\,V^{1:2}_{\alpha\eta\omega}(\gamma^*)]$$

where $U_{33} = \frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq s}\epsilon_i\Delta_j\mathbf{Z}_{ki}^T(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I)\mathbf{Z}_{kj}$. Since the above holds for all r, a

with finite variance and a having mean 0, we have

$$
\begin{aligned}
& \left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\mathbf{Z}_{kj} \\
= \quad & \frac{1}{(n-1)(n-2)}\left(\frac{n-1}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right)^2 \\
& \left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\mathbb{E}\left[\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] \\
& -\frac{1}{(n-2)}\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}\mathbb{E}\left[\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] \\
& -\frac{1}{(n-2)}\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}\mathbb{E}\left[\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] \\
& +\frac{1}{(n-2)}\left\{\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}_s\right) - \mathbb{E}\left[\mathbf{Z}_k\mathbf{Z}_k^T\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right]\right\} \\
& +\frac{1}{(n-2)}\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}^2\mathbb{E}\left[\mathbf{Z}_k\mathbf{Z}_k^T\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] \\
& +\frac{1}{(n-2)}\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}_s\right) + \gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}_s\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\mathbf{Z}_{kj} \\
= \quad & \frac{1}{(n-1)(n-2)}\left(\frac{n-1}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right)^2 \\
& \times\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\mathbb{E}\left[\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] \\
& -\frac{2}{(n-2)}\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\} \\
& \times\mathbb{E}\left[\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] \\
& +\frac{1}{(n-2)}\left\{\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}_s\right) - \mathbb{E}\left[\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right]\right\} \\
& +\frac{1}{(n-2)}\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^T - I\right)\left\{I - \frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T}{(n-1)+\mathbf{Z}_{kj}^T\mathbf{Z}_{kj}}\right\}\mathbb{E}\left[\left\{\mathbf{Z}_k\mathbf{Z}_k^T - I\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] \\
& +\left\{I + \frac{1}{(n-2)}\mathbf{Z}_{kj}\mathbf{Z}_{kj}^T\right\}\gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}_s\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& \left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)\mathbf{Z}_{kj} \\
= \quad & \frac{1}{(n-1)(n-2)}\left(\frac{n-1}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right)^{2} \\
& \times\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right] \\
& -\frac{2}{(n-2)}\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)\left\{I-\frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right\} \\
& \times\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right] \\
& -\frac{1}{(n-2)}\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right] \\
& +\frac{\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)}{(n-2)}\left\{I-\frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right\}^{2}\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right] \\
& +\left\{I+\frac{1}{(n-2)}\left\{\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}+\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}\right\}\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}_{s}\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \left\{I+\frac{1}{(n-2)}\left\{\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}+\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}\right\}\right\}^{-1}\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)\mathbf{Z}_{kj} \\
= \quad & \left\{I+\frac{\left\{\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}+\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}\right\}}{(n-2)}\right\}^{-1}\{\ \frac{1}{(n-1)(n-2)}\left(\frac{n-1}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right)^{2} \\
& \times\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right] \\
& -\frac{2\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)}{(n-2)}\left\{I-\frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right\}\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right] \\
& -\frac{1}{(n-2)}\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right] \\
& +\frac{\left(\mathbf{Z}_{ks}\mathbf{Z}_{ks}^{T}-I\right)}{(n-2)}\left\{I-\frac{\mathbf{Z}_{kj}\mathbf{Z}_{kj}^{T}}{(n-1)+\mathbf{Z}_{kj}^{T}\mathbf{Z}_{kj}}\right\}^{2}\mathbb{E}\left[\left\{\mathbf{Z}_{k}\mathbf{Z}_{k}^{T}-I\right\}\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}\right)\right]\ \} \\
& +\gamma^{*}\left(\boldsymbol{X}_{j},\boldsymbol{X}_{s}\right).
\end{aligned}
$$

Finally, with $\mathbb{E}_s$ denoting expectation with respect to $\boldsymbol{X}_s$, we have

$$
\begin{aligned}
& \mathbb{E}_s \left[\mathbf{Z}_{ks} \mathbf{Z}_{ks}^T \left\{ I + \frac{1}{(n-2)} \left\{ \mathbf{Z}_{ks} \mathbf{Z}_{ks}^T + \mathbf{Z}_{kj} \mathbf{Z}_{kj}^T \right\} \right\}^{-1} \left(\mathbf{Z}_{ks} \mathbf{Z}_{ks}^T - I \right) \mathbf{Z}_{kj} \right] \\
= \; & \mathbb{E}_s \left[\, \mathbf{Z}_{ks} \mathbf{Z}_{ks}^T \left\{ I + \frac{1}{(n-2)} \left\{ \mathbf{Z}_{ks} \mathbf{Z}_{ks}^T + \mathbf{Z}_{kj} \mathbf{Z}_{kj}^T \right\} \right\}^{-1} \times \right. \\
& \left\{ \frac{1}{(n-1)(n-2)} \left(\frac{n-1}{(n-1) + \mathbf{Z}_{kj}^T \mathbf{Z}_{kj}} \right)^2 \right. \\
& \times \left(\mathbf{Z}_{ks} \mathbf{Z}_{ks}^T - I \right) \mathbf{Z}_{kj} \mathbf{Z}_{kj}^T \mathbb{E} \left[\left\{ \mathbf{Z}_k \mathbf{Z}_k^T - I \right\} \gamma^* \left(\boldsymbol{X}_j, \boldsymbol{X} \right) \right] \\
& - \frac{2}{(n-2)} \left(\mathbf{Z}_{ks} \mathbf{Z}_{ks}^T - I \right) \left\{ I - \frac{\mathbf{Z}_{kj} \mathbf{Z}_{kj}^T}{(n-1) + \mathbf{Z}_{kj}^T \mathbf{Z}_{kj}} \right\} \\
& \mathbb{E} \left[\left\{ \mathbf{Z}_k \mathbf{Z}_k^T - I \right\} \gamma^* \left(\boldsymbol{X}_j, \boldsymbol{X} \right) \right] \\
& - \frac{1}{(n-2)} \mathbb{E} \left[\left\{ \mathbf{Z}_k \mathbf{Z}_k^T - I \right\} \gamma^* \left(\boldsymbol{X}_j, \boldsymbol{X} \right) \right] \\
& + \frac{\left(\mathbf{Z}_{ks} \mathbf{Z}_{ks}^T - I \right)}{(n-2)} \left\{ I - \frac{\mathbf{Z}_{kj} \mathbf{Z}_{kj}^T}{(n-1) + \mathbf{Z}_{kj}^T \mathbf{Z}_{kj}} \right\} \mathbb{E} \left[\left\{ \mathbf{Z}_k \mathbf{Z}_k^T - I \right\} \gamma^* \left(\boldsymbol{X}_j, \boldsymbol{X} \right) \right] \left. \left. \right\} \right] \\
& + \mathbb{E} \left[\mathbf{Z}_k \mathbf{Z}_k^T \gamma^* \left(\boldsymbol{X}_j, \boldsymbol{X} \right) \right]
\end{aligned}
$$

Now, solving for $\mathbb{E}\left[\left\{\mathbf{Z}_k \mathbf{Z}_k^T - I\right\} \gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right] = \mathbb{E}\left[\mathbf{Z}_k \mathbf{Z}_k^T \gamma^*\left(\boldsymbol{X}_j, \boldsymbol{X}\right)\right]$ and plugging the solution into the earlier expression ends the proof.

□